Convolutional Nets
Part 2

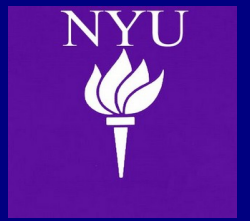
Yann Le Cun

Facebook AI Research,

Center for Data Science, NYU

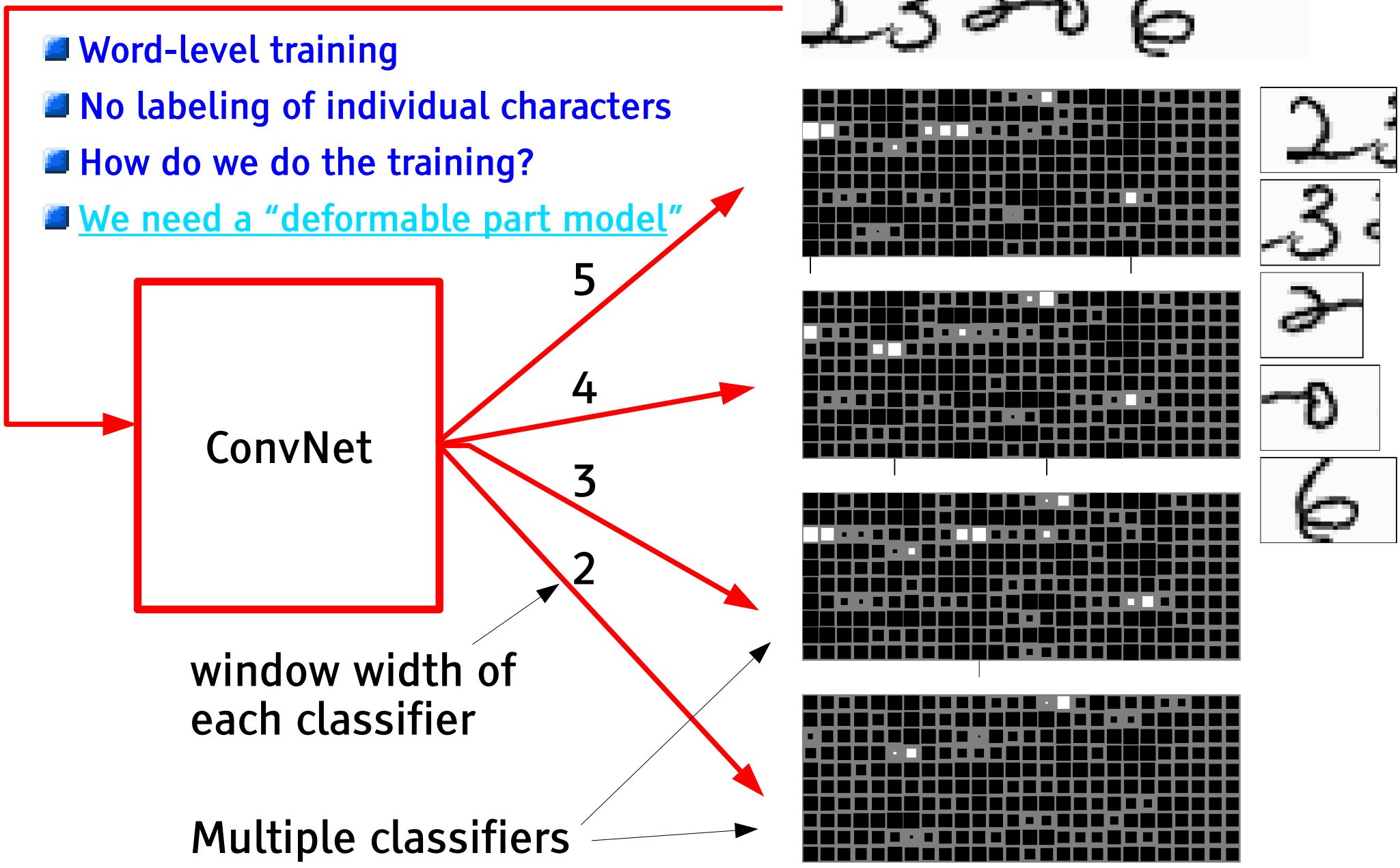
Courant Institute of Mathematical Sciences, NYU

http://yann.lecun.com



Word-level training with weak supervision [Matan et al 1992]

- Word-level training
- No labeling of individual characters
- How do we do the training?
- We need a "deformable part model"



Graph Transformer Networks

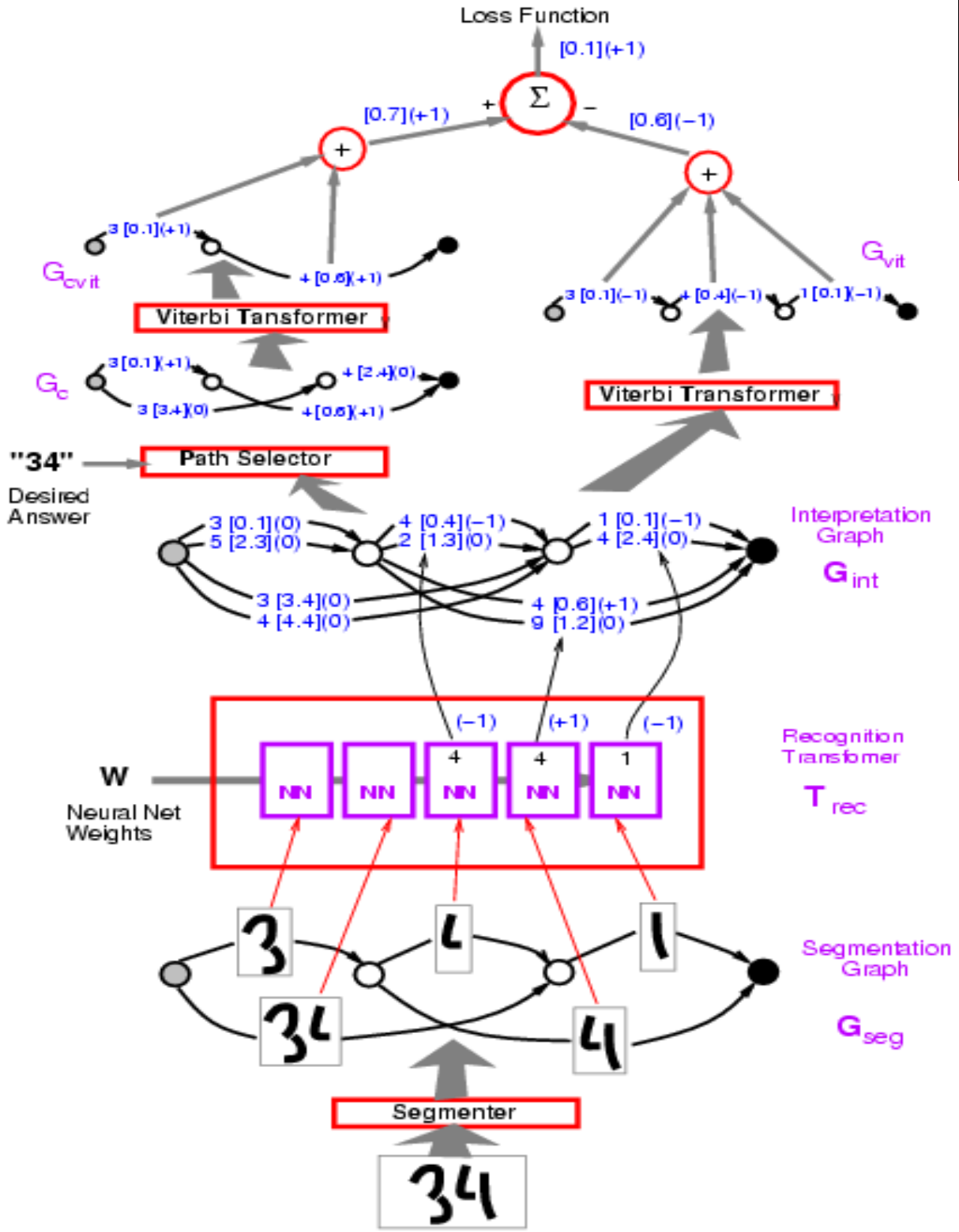
Structured Prediction

on top of Deep Learning

This example shows the structured perceptron loss.

In practice, we used negative log-likelihood loss.

Deployed in 1996 in check reading machines.



Check Reader (Bell Labs, 1995)

Graph transformer network trained to read check amounts.

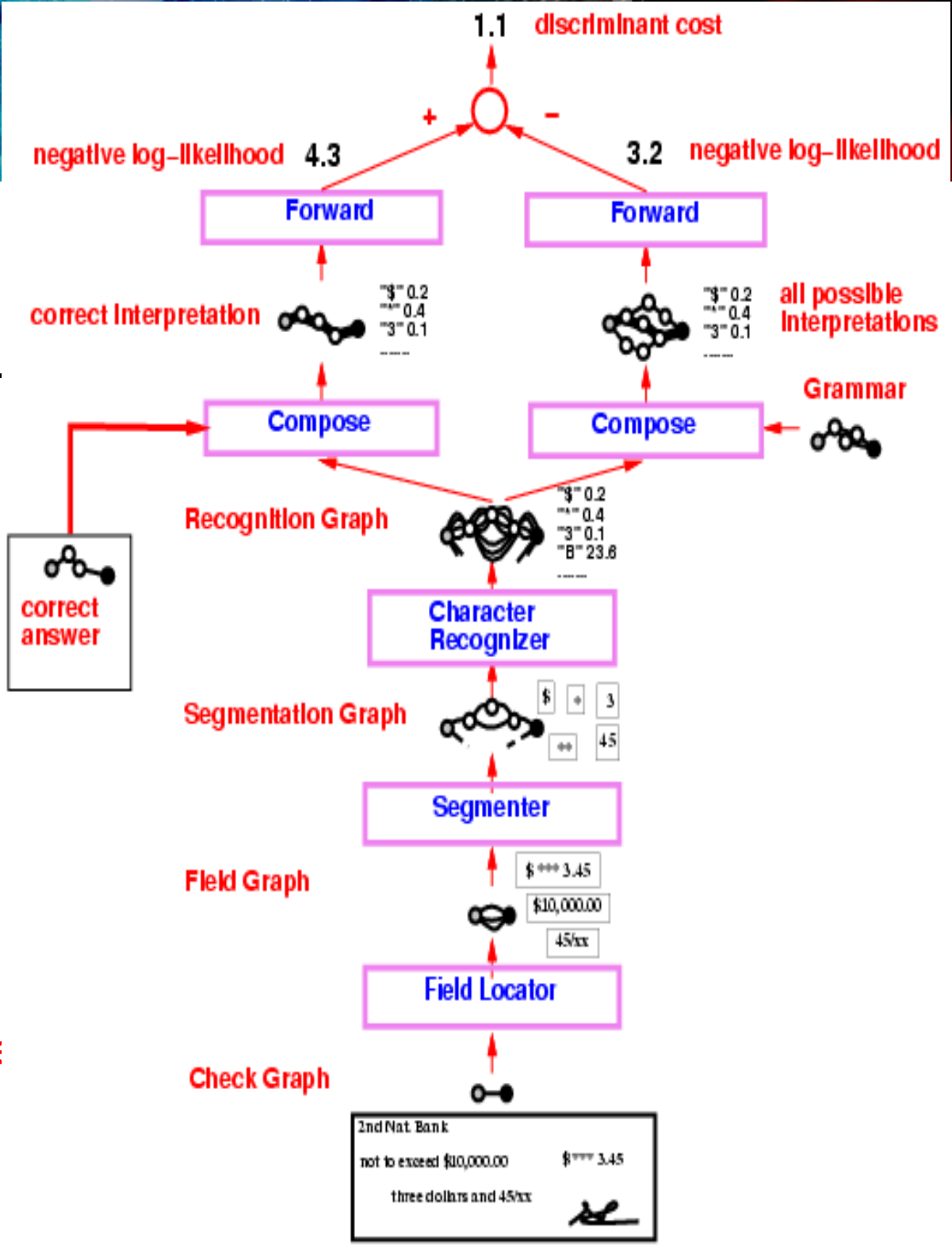
Trained globally with Negative-Log-Likelihood loss.

50% percent correct, 49% reject, 1% error (detectable later in the process).

Fielded in 1996, used in many banks in the US and Europe.

Processed an estimated 10% to 20% of all the checks written in the US in the early 2000s.

[LeCun, Bottou, Bengio, Haffner 1998]

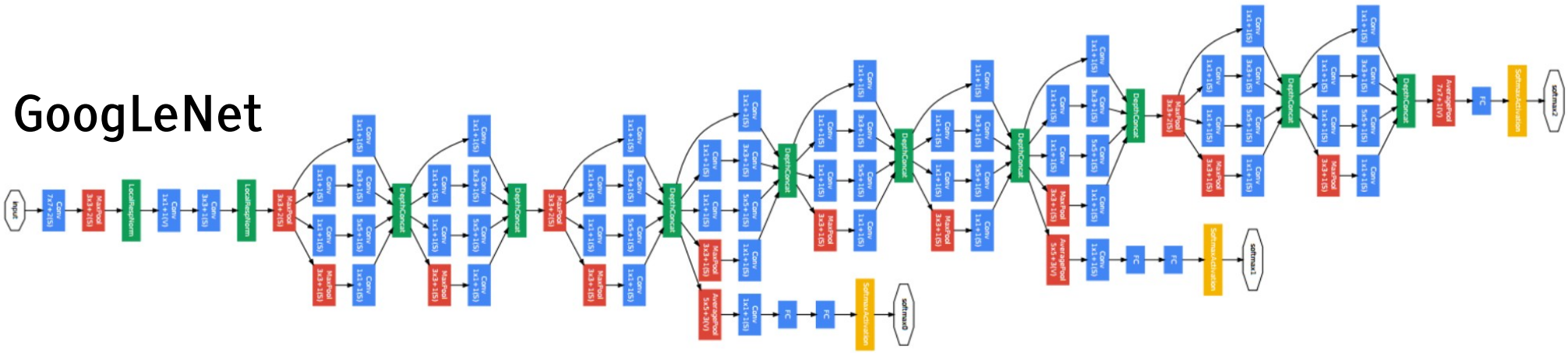


Very Deep ConvNet Architectures

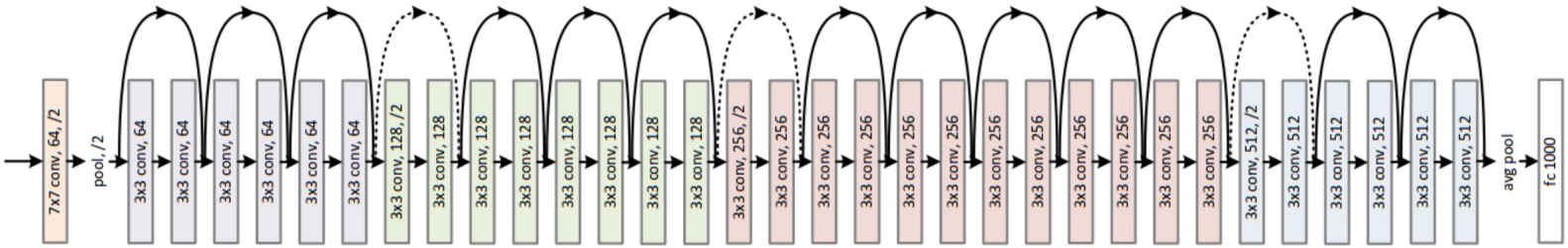
- Small kernels, not much subsampling (fractional subsampling).

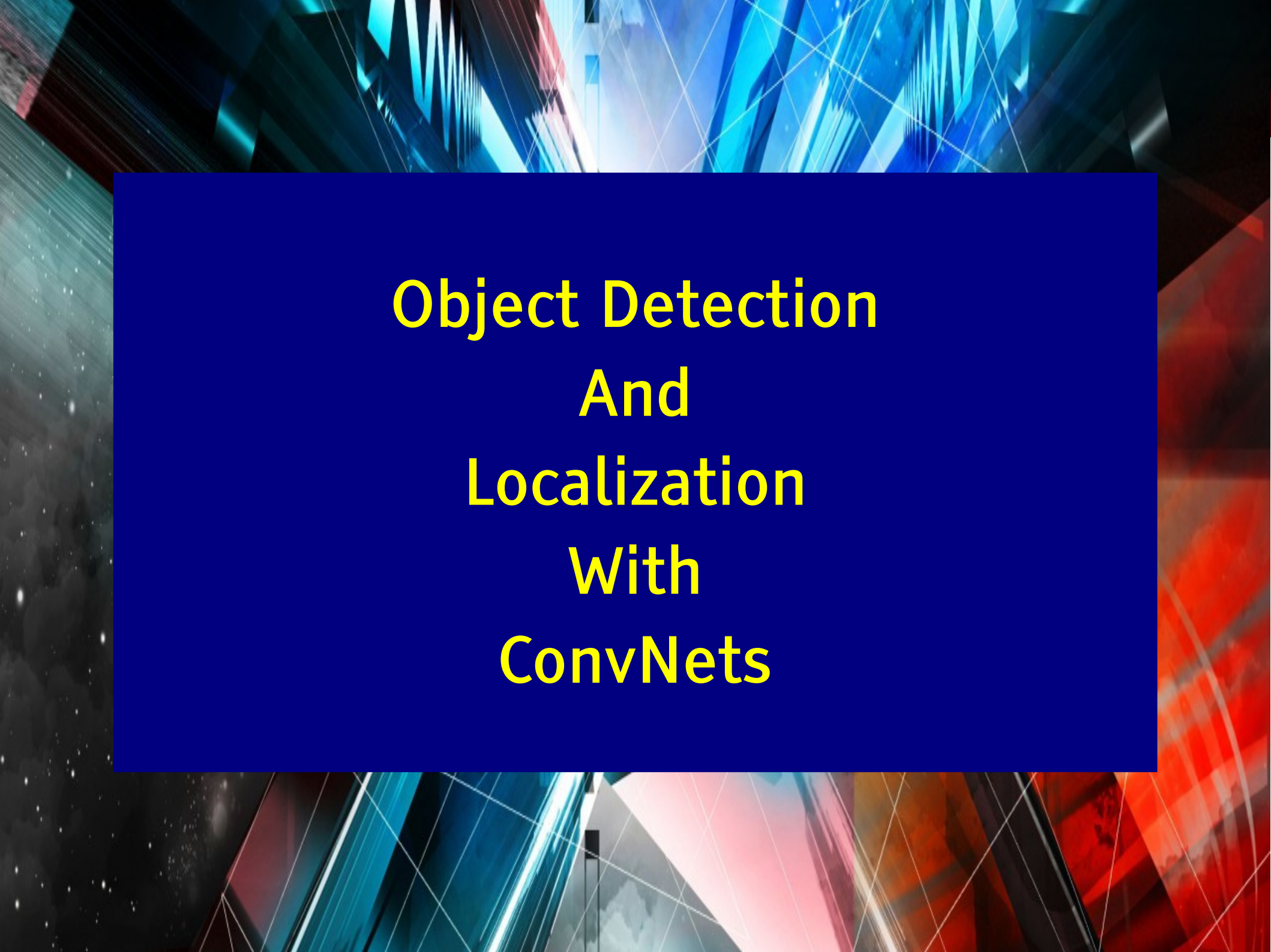
VGG

GoogLeNet



ResNet




Object Detection
And
Localization
With
ConvNets

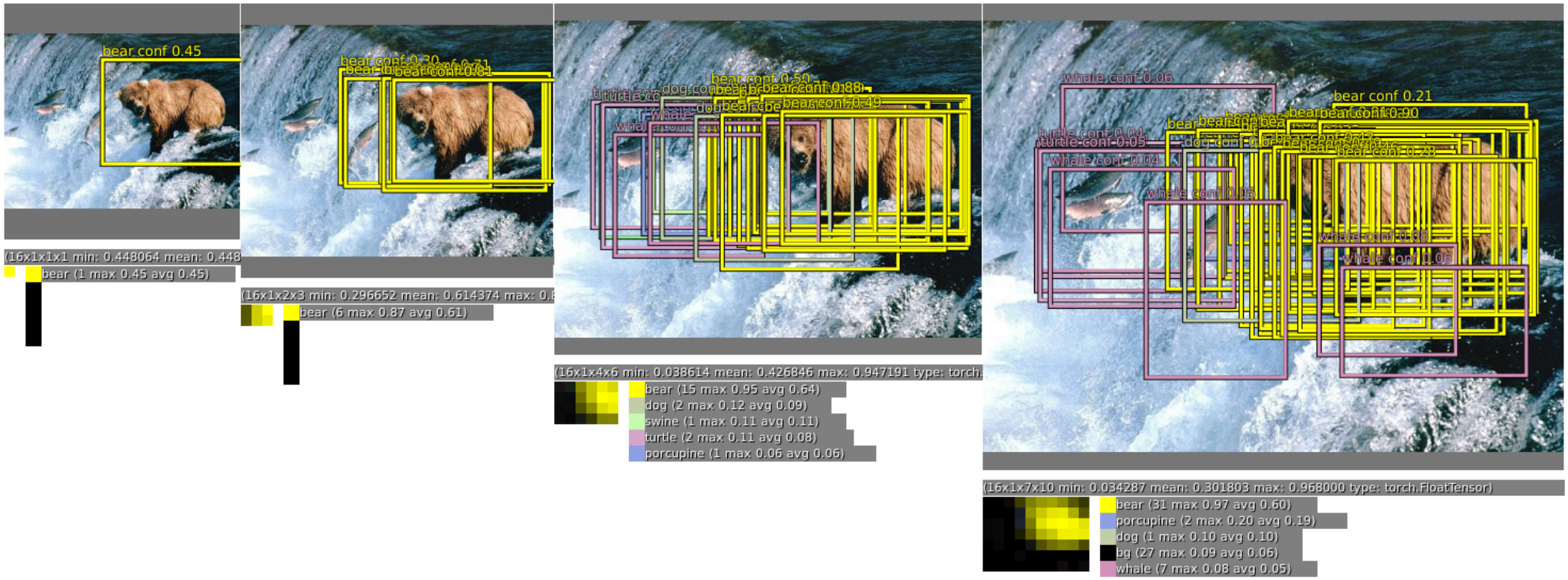
Classification + Localization: multiscale sliding window

- Apply convnet with a sliding window over the image at multiple scales
- Important note: it's very cheap to slide a convnet over an image
  - Just compute the convolutions over the whole image and replicate the fully-connected layers

Classification + Localization: sliding window + bounding box regression

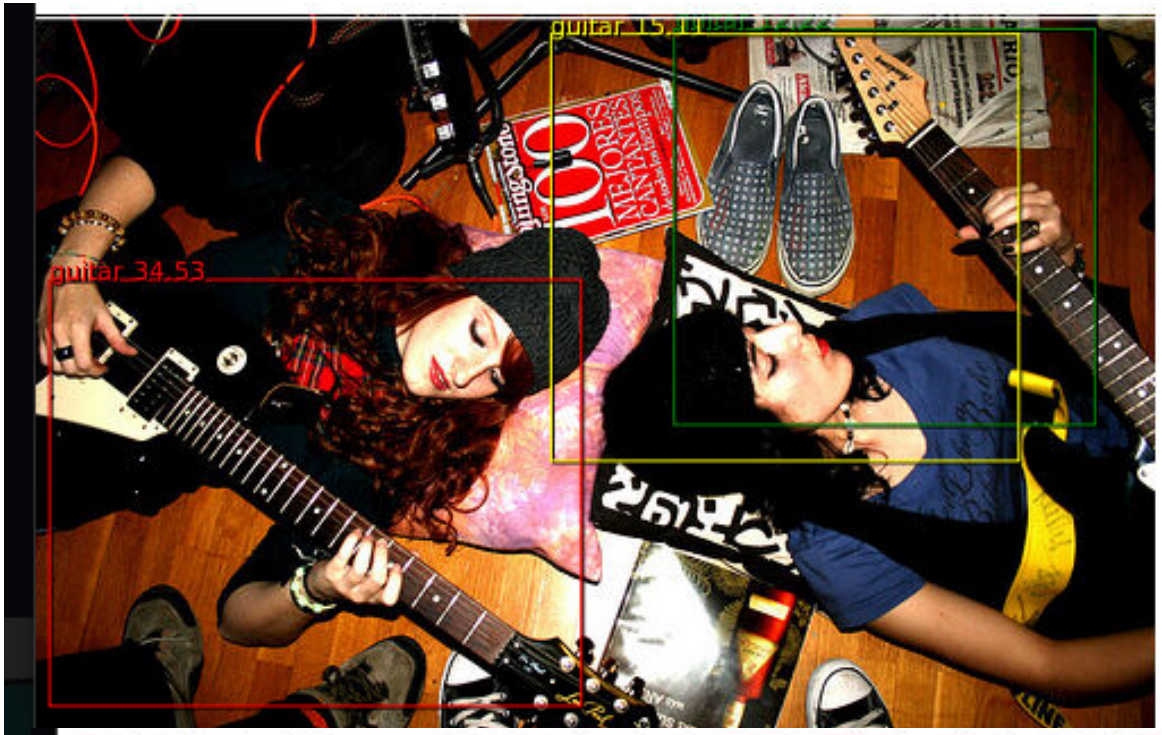
Y LeCun

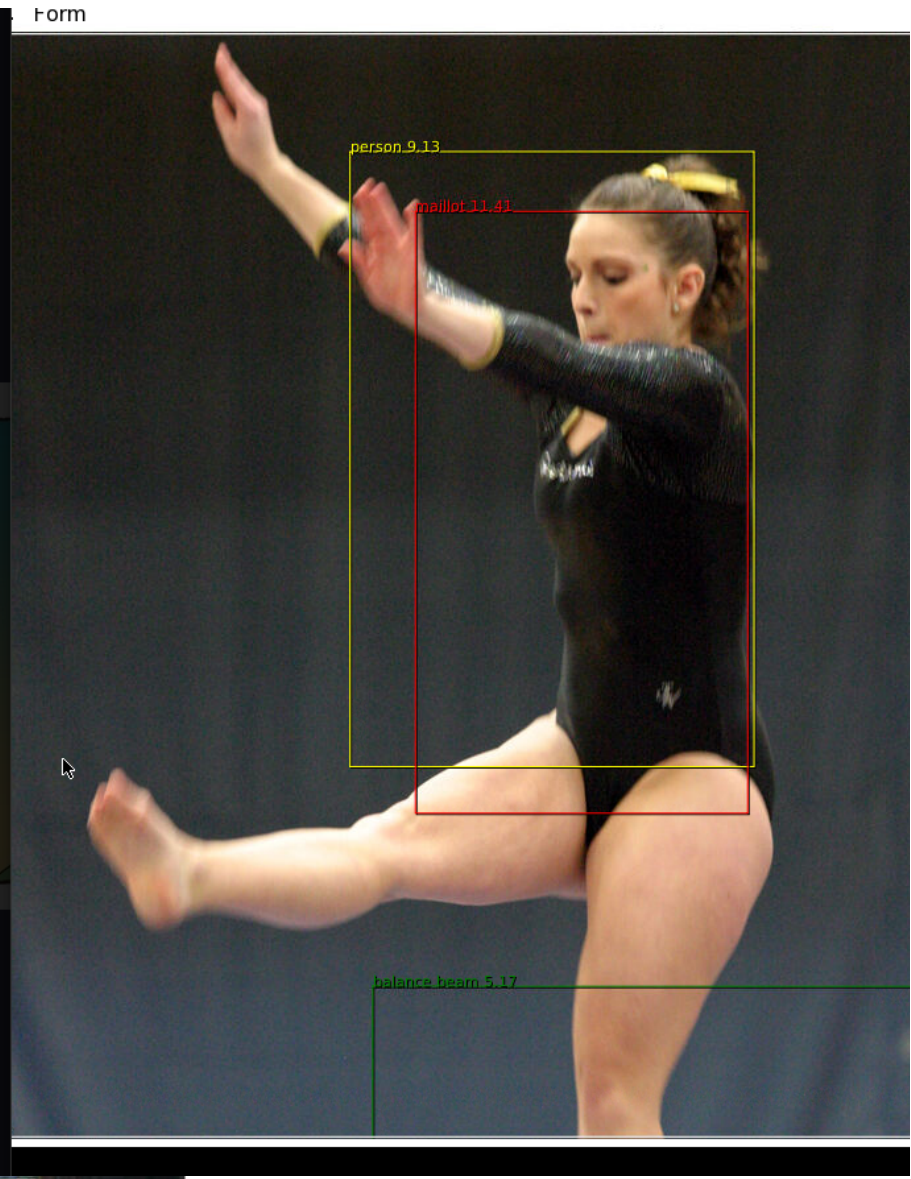
- Apply convnet with a sliding window over the image at multiple scales
- For each window, predict a class and bounding box parameters
  - Evenif the object is not completely contained in the viewing window, the convnet can predict where it thinks the object is.

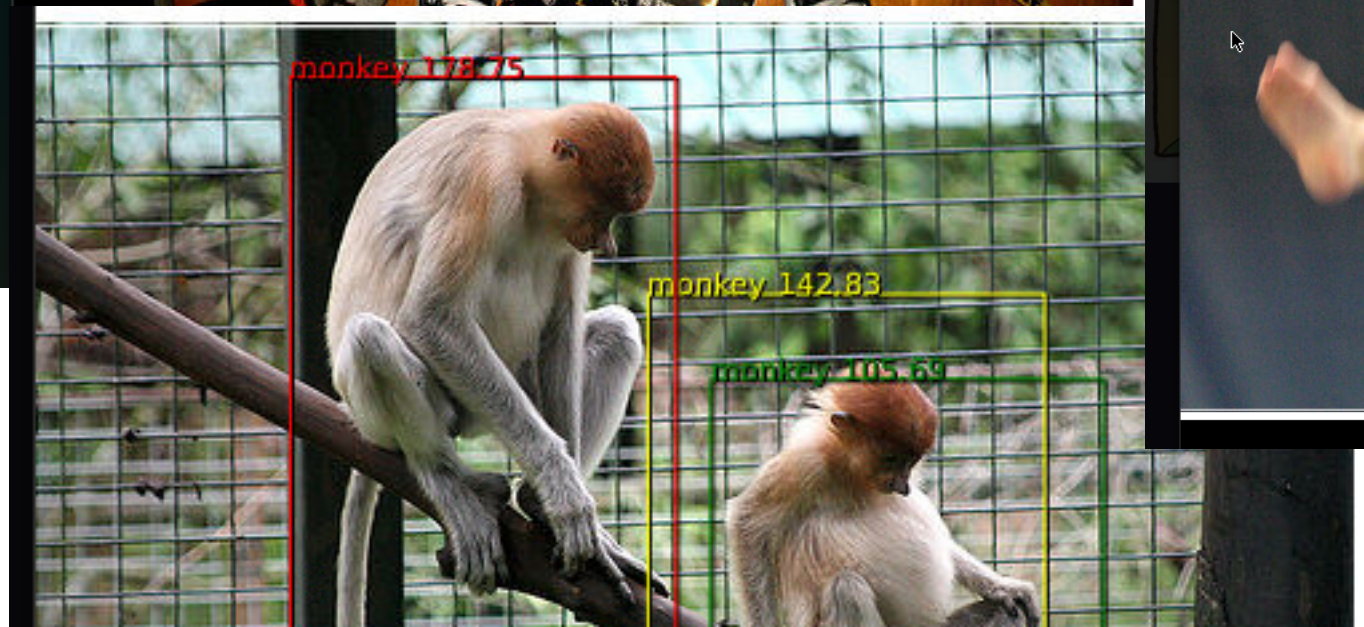


Results: pre-trained on ImageNet1K, fine-tuned on ImageNet Detection

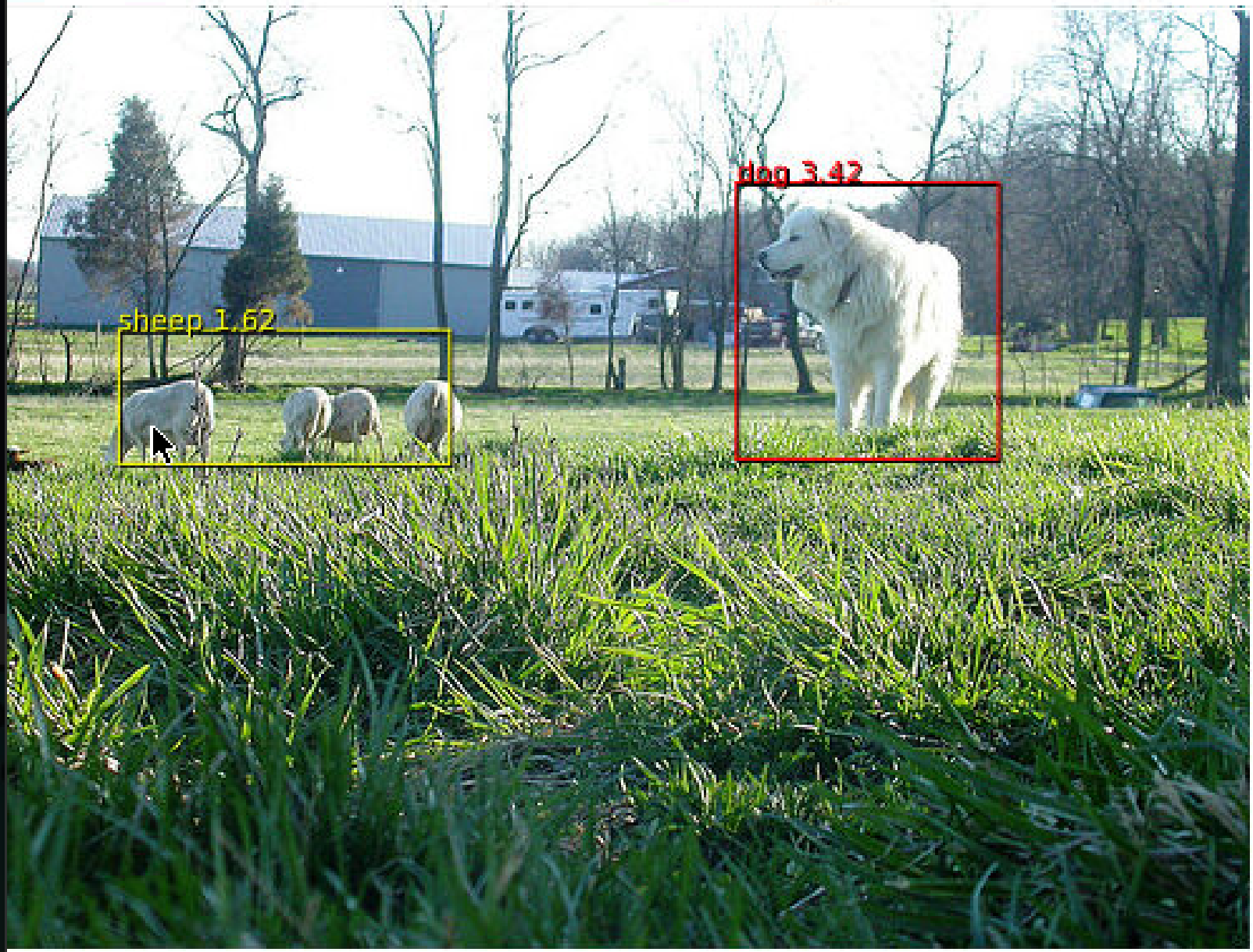
Y LeCun





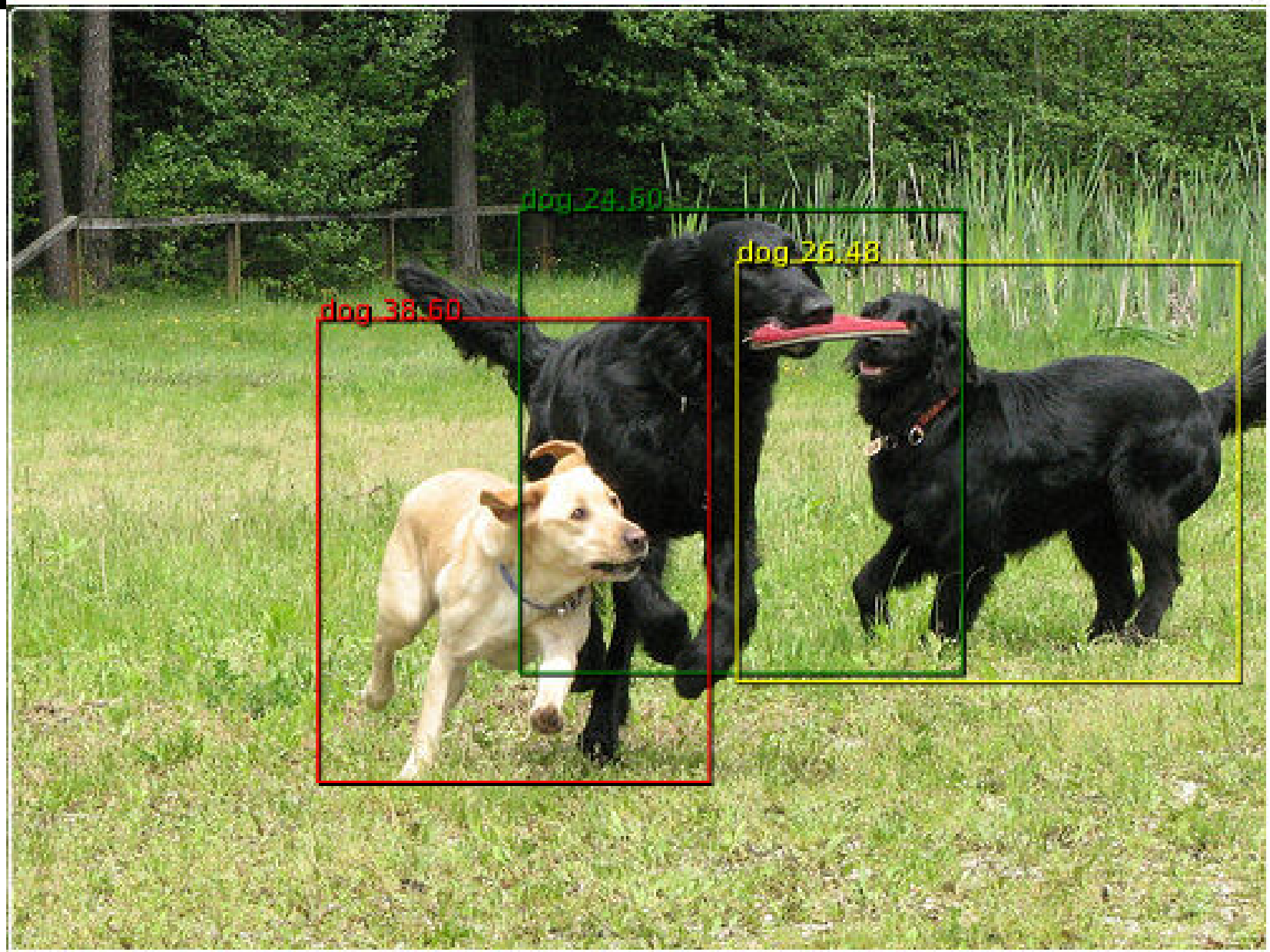


Detection Examples



/home/snwiz/data/imagenet12/original/det/ILSVRC2013_DET_test/ILSVRC2012_test_00090628.JPEG
dog conf 3.419652

Detection Examples

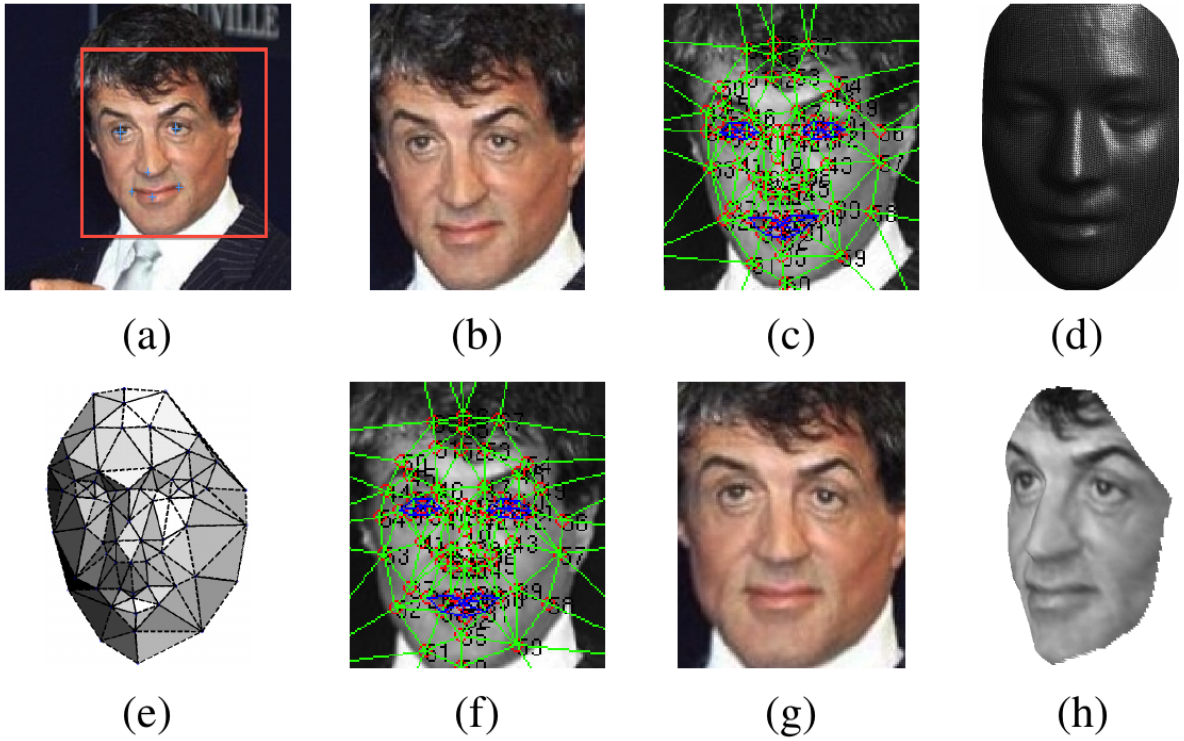


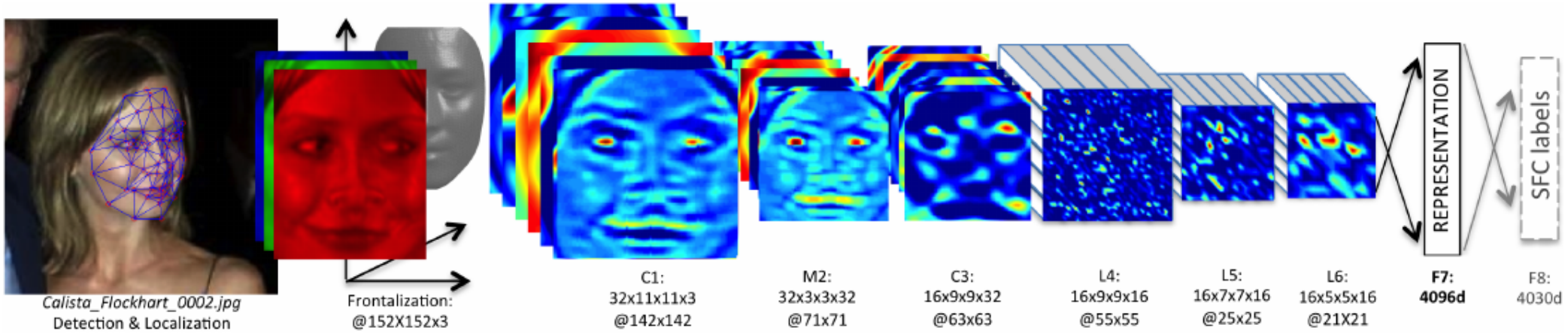
/home/snwiz/data/imagenet12/original/det/ILSVRC2013_DET_test/ILSVRC2012_test_00000172.JPEG
dog conf 38.603936

Detection Examples

Deep Face

- **[Taigman et al. CVPR 2014]**
  - Alignment
  - ConvNet
  - Metric Learning
- **Deployed at Facebook for Auto-tagging**
  - 800 million photos per day

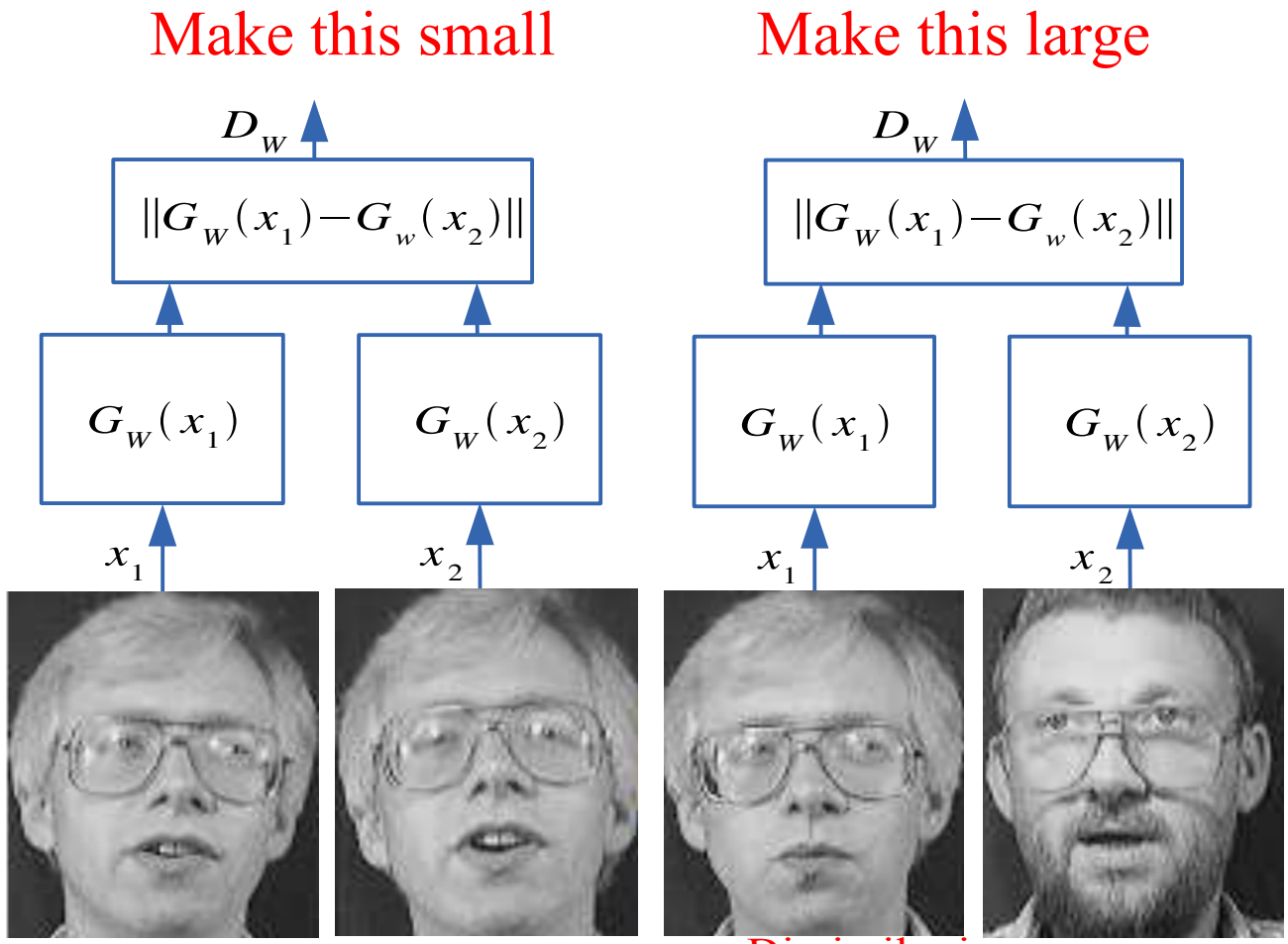




Metric Learning with a Siamese Architecture

Contrative Obective Function

- Similar objects should produce outputs that are nearby
- Dissimilar objects should produce output that are far apart.
- DrLIM: Dimensionality Reduction by Learning and Invariant Mapping
- [Chopra et al. CVPR 2005]
- [Hadsell et al. CVPR 2006]



Pose Estimation and Attribute Recovery with ConvNets

Y LeCun

Pose-Aligned Network for Deep Attribute Modeling

[Zhang et al. CVPR 2014] (Facebook AI Research)

(a) Highest scoring results for people wearing glasses.

(b) Highest scoring results for people wearing a hat.

Real-time hand pose recovery

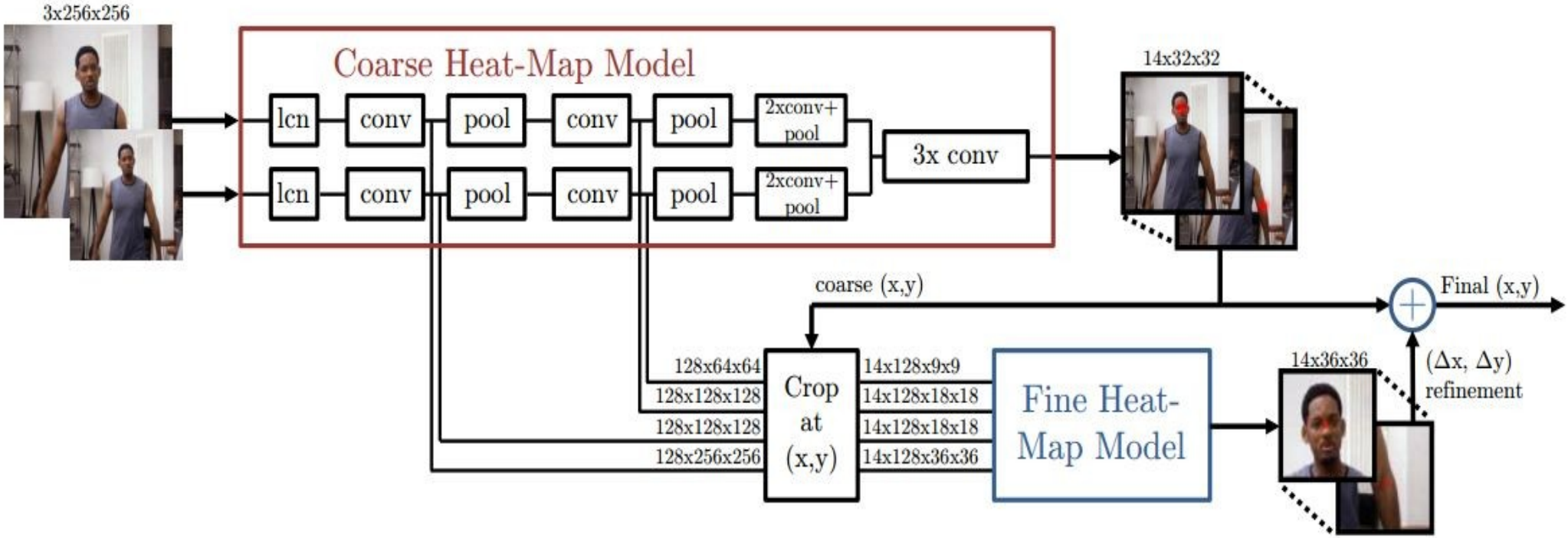
[Tompson et al. Trans. on Graphics 14]

Body pose estimation [Tompson et al. ICLR, 2014]

HAND POSE V IDEO

Person Detection and Pose Estimation

Tompson, Goroshin, Jain, LeCun, Bregler arXiv:1411.4280 (2014)



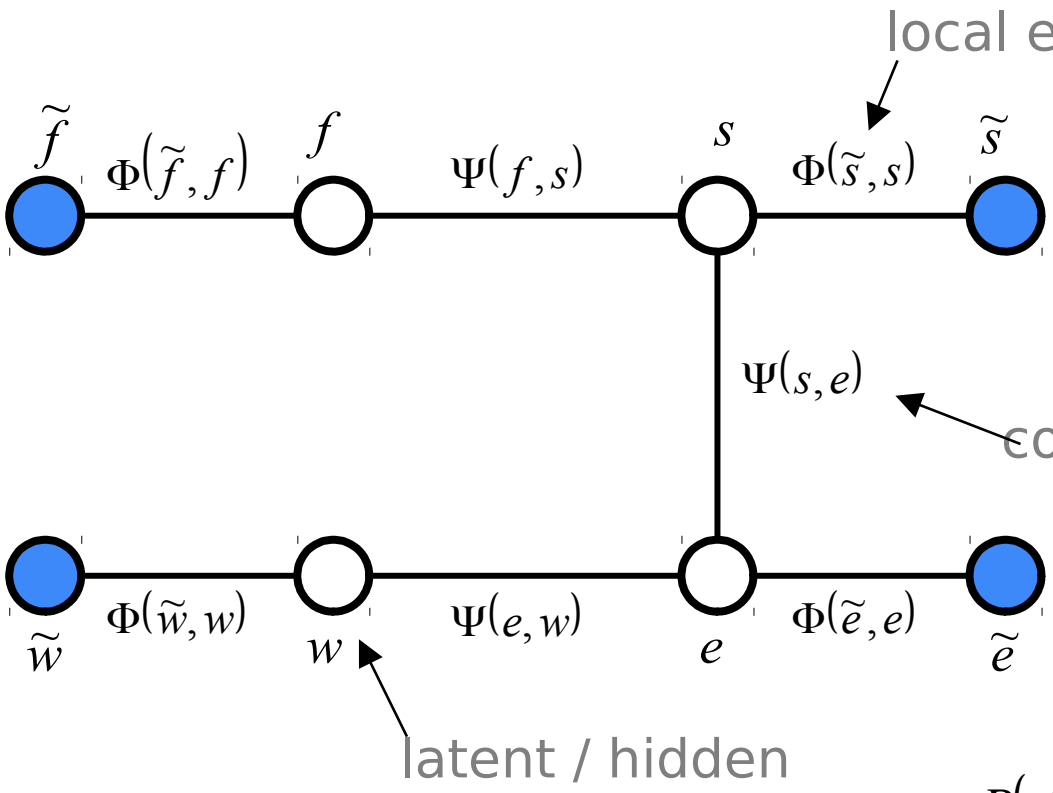
Person Detection and Pose Estimation

Tompson, Goroshin, Jain, LeCun, Bregler arXiv:1411.4280 (2014)

SPATIAL MODEL

Start with a tree graphical model

MRF over spatial locations



Joint Distribution:

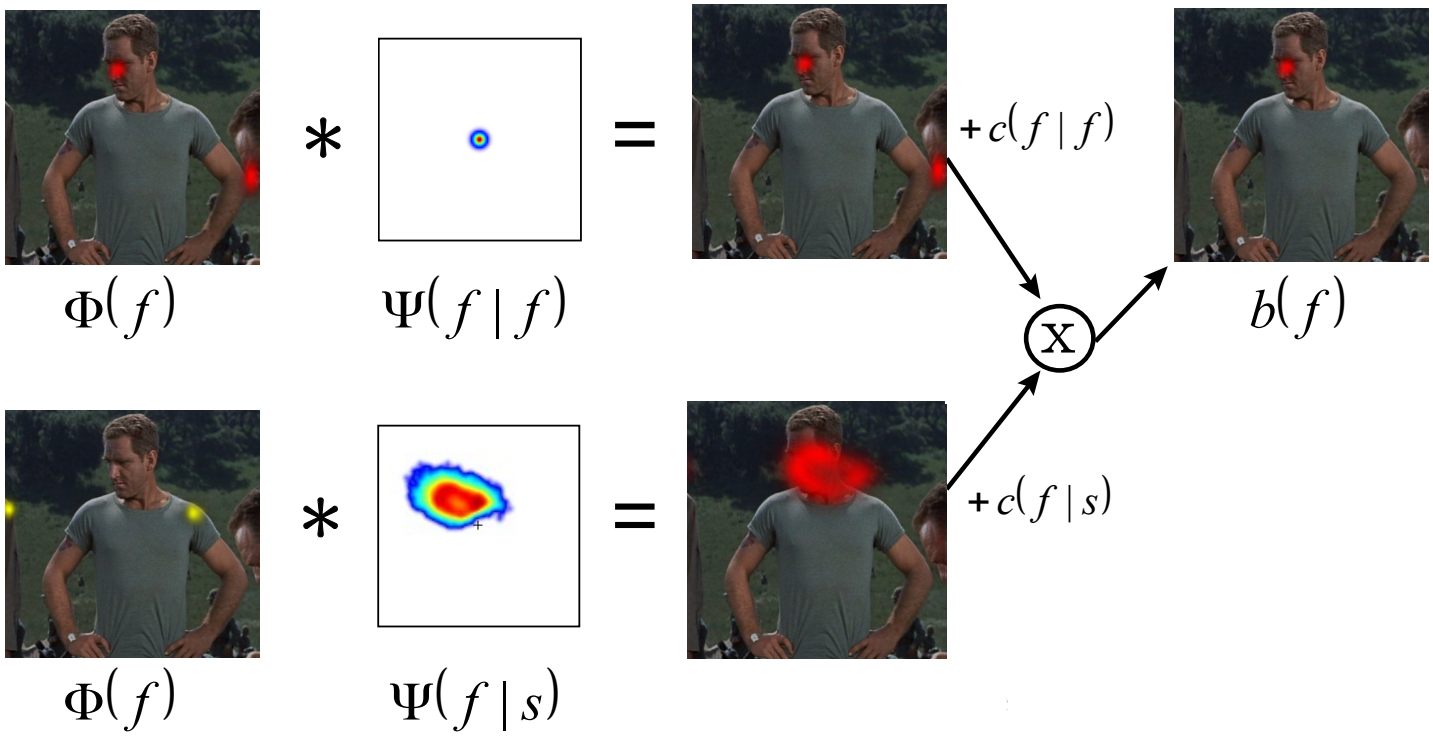
$$P(f,s,e,w) = \frac{1}{Z} \prod_{i,j} \Psi(x_i, x_j) \prod_i \Phi(x_i, \tilde{x}_i)$$

SPATIAL MODEL

Start with a tree graphical model

... And approximate it

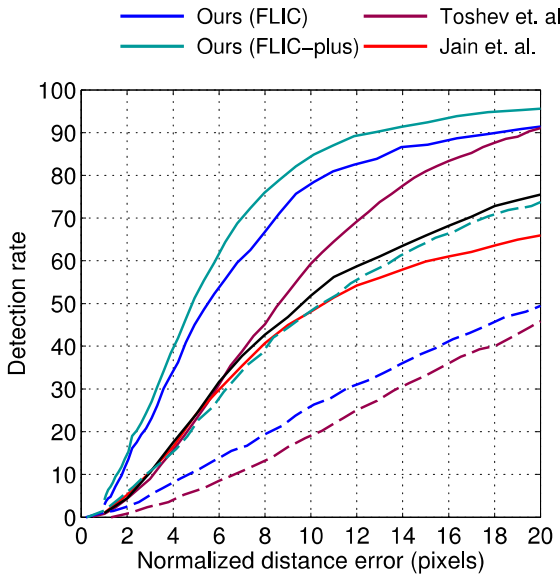
$$b(f) = \Phi(f) \prod_i (\Phi(x_i) \star \Psi(f \mid x_i) + c(f \mid x_i))$$



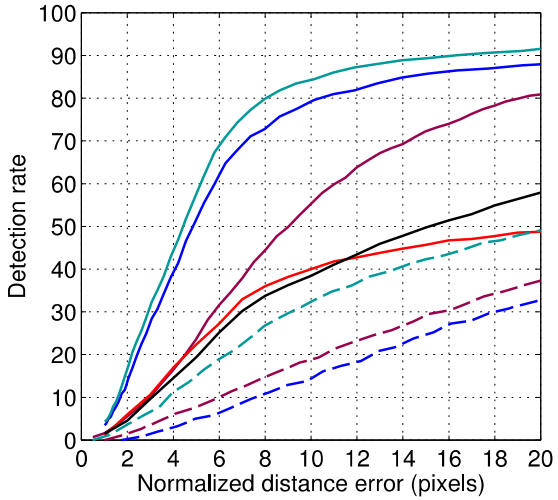
SPATIAL MODEL: RESULTS

Y LeCun

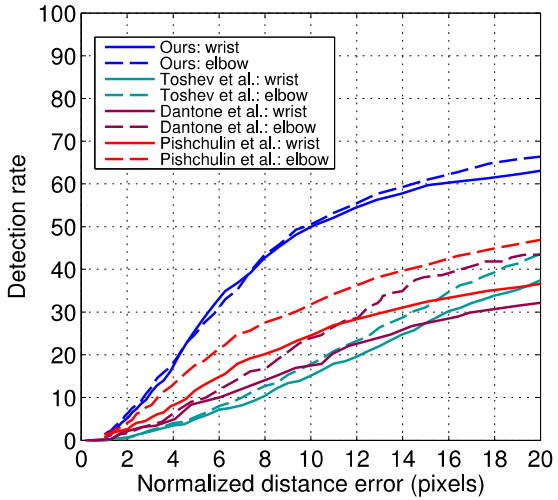
FLIC[1] Elbow



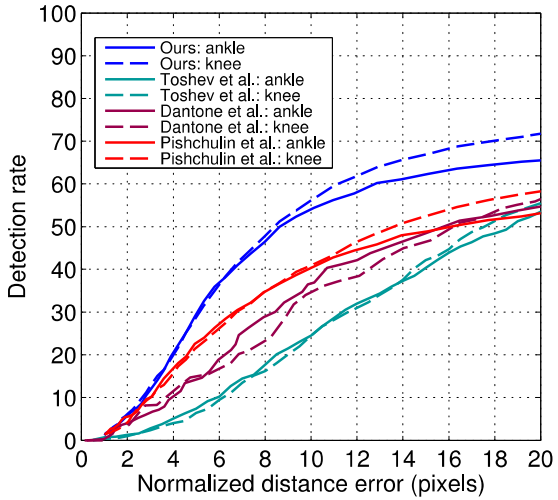
FLIC[1] Wrist



LSP[2] Arms



LSP[1] Legs



(1) B. Sapp and B. Taskar. MODEC: Multimodel decomposition models for human pose estimation. CVPR'13
(2) S. Johnson and M. Everingham. Learning Effective Human Pose Estimation for Inaccurate Annotation. CVPR'11

Image captioning: generating a descriptive sentence

Y LeCun

[Lebret, Pinheiro, Collobert 2015] [Kulkarni 11] [Mitchell 12] [Vinyals 14] [Mao 14]

A man riding skis on a snow covered ski slope.

NP: a man, skis, the snow, a person, a woman, a snow covered slope, a slope, a snowboard, a skier, man.
VP: wearing, riding, holding, standing on, skiing down.
PP: on, in, of, with, down.

A man wearing skis on the snow.

A man is doing skateboard tricks on a ramp.

NP: a skateboard, a man, a trick, his skateboard, the air, a skateboarder, a ramp, a skate board, a person, a woman.
VP: doing, riding, is doing, performing, flying through.
PP: on, of, in, at, with.

A man riding a skateboard on a ramp.

The girl with blue hair stands under the umbrella.

NP: a woman, an umbrella, a man, a person, a girl, umbrellas, that, a little girl, a cell phone.
VP: holding, wearing, is holding, holds, carrying.
PP: with, on, of, in, under.

A woman is holding an umbrella.

A slice of pizza sitting on top of a white plate.

NP: a plate, a white plate, a table, pizza, it, a pizza, food, a sandwich, top, a close.
VP: topped with, has, is, sitting on, is on.
PP: of, on, with, in, up.

A table with a plate of pizza on a white plate.

A baseball player swinging a bat on a field.

NP: the ball, a game, a baseball player, a man, a tennis court, a ball, home plate, a baseball game, a batter, a field.
VP: swinging, to hit, playing, holding, is swinging.
PP: on, during, in, at, of.

A baseball player swinging a bat on a baseball field.

A bunch of kites flying in the sky on the beach.

NP: the beach, a beach, a kite, kites, the ocean, the water, the sky, people, a sandy beach, a group.
VP: flying, flies, is flying, flying in, are.
PP: on, of, with, in, at.

People flying kites on the beach.

C3D: Video Classification with 3D ConvNet

[Tran et al. 2015]

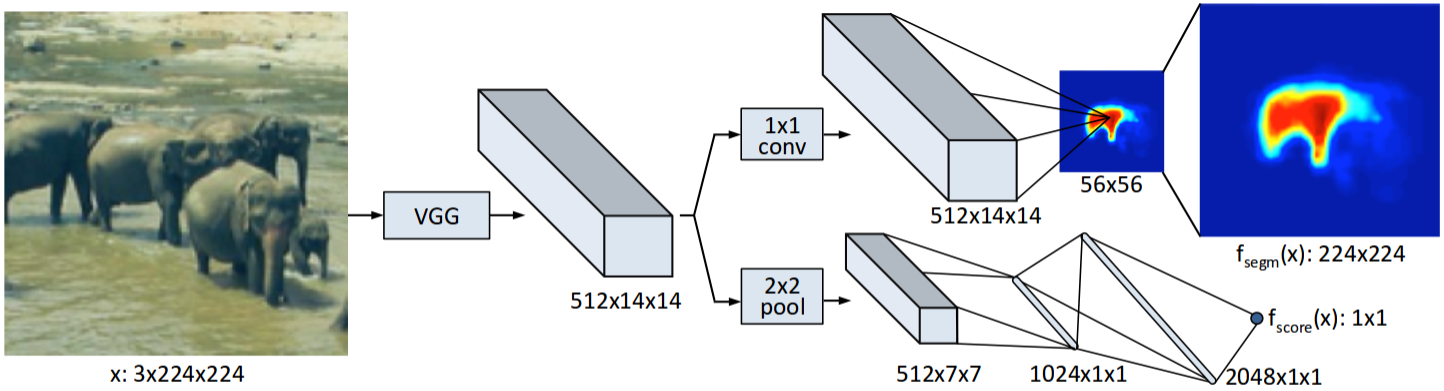
VIDEO: COMMON SPORTS

VIDEO: UNCOMMON SPORTS

Segmenting and Localizing Objects (DeepMask)

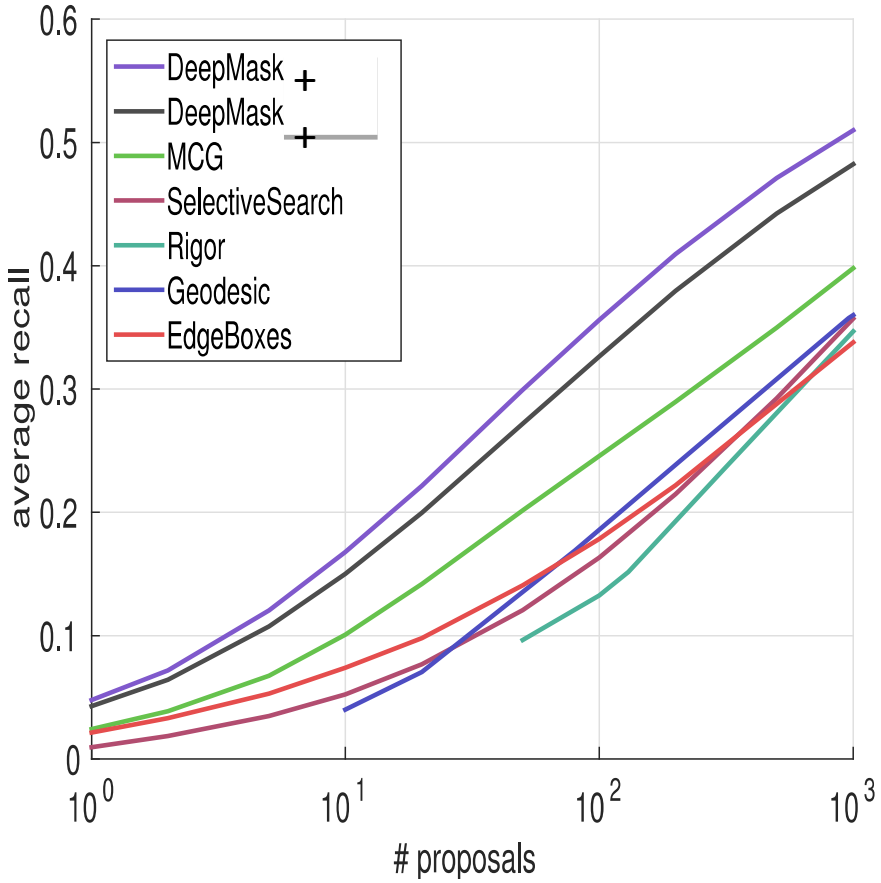
Y LeCun

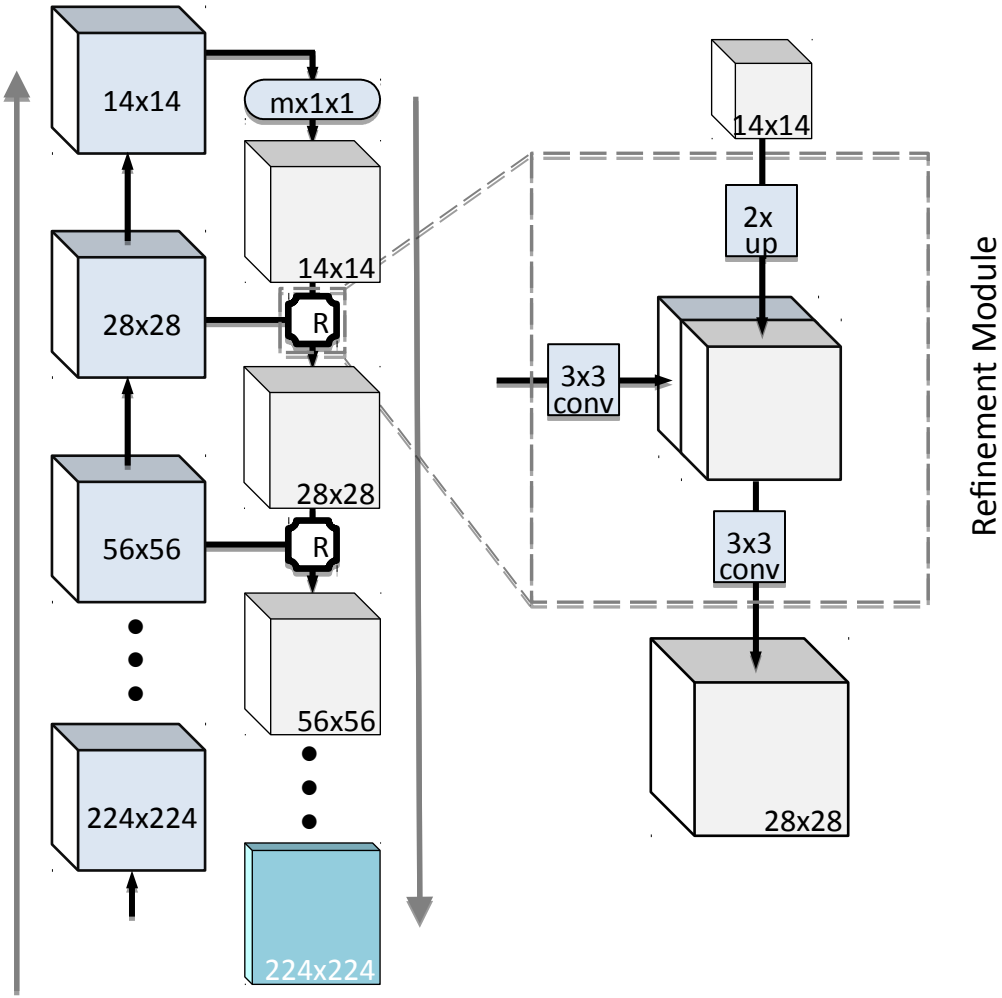
- [Pinheiro, Collobert, Dollar ICCV 2015]
  - ConvNet produces object masks



Iterative Refinement: DeepMask++

FAIR COCO Team



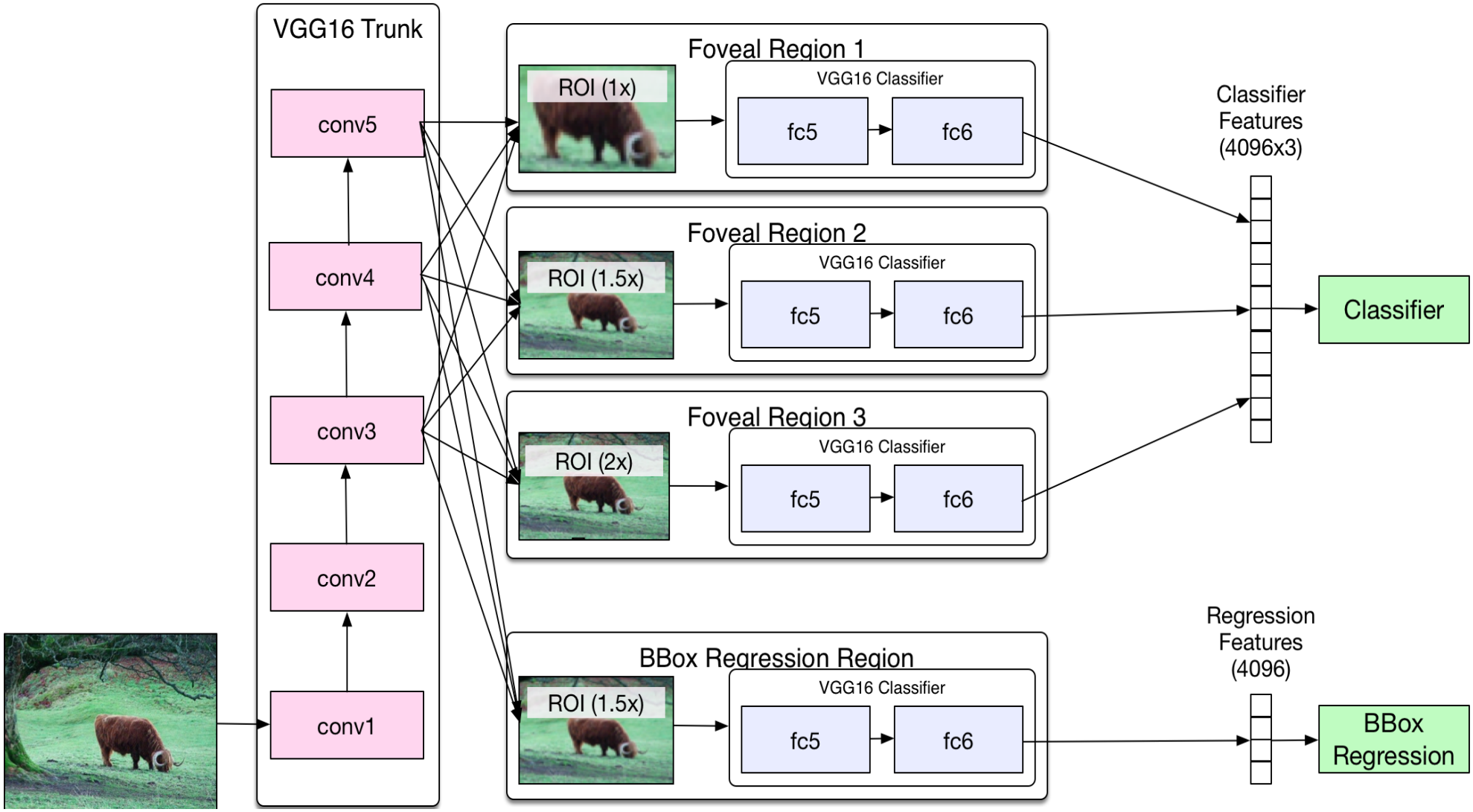


DeepMask++ Proposals

FAIR COCO Team

Recognition Pipeline

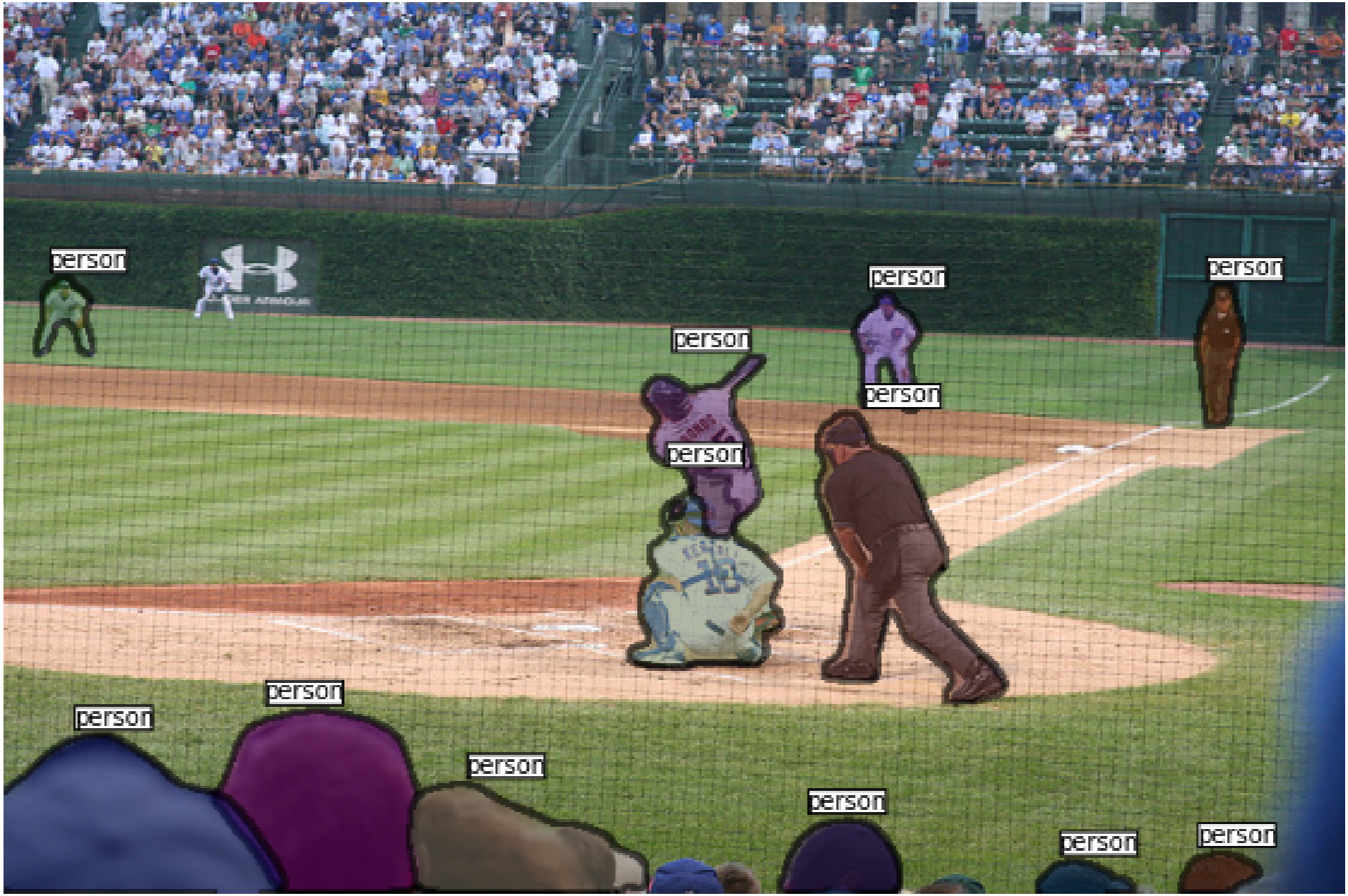
FAIR COCO Team



- 2.5 days on 8x4 Kepler GPUs with Elastic Avergaing Stochastic Gradient Descent (EASGD [Zhang, Choromanska, LeCun NIPS 2015]
- "Big Sur"

Results

Results


person
person
person
person
person
person
person
person
person
person
person
person

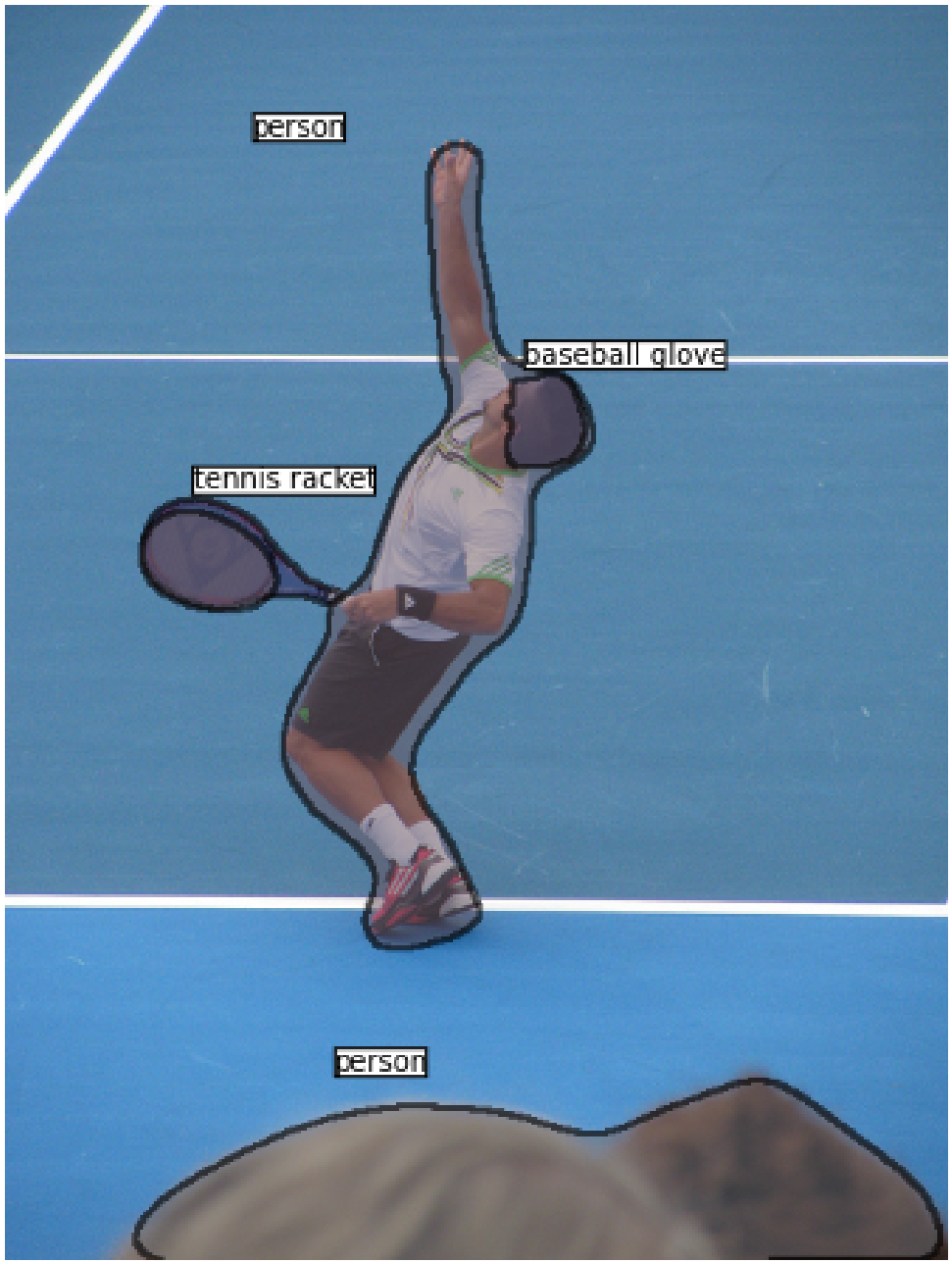
Results

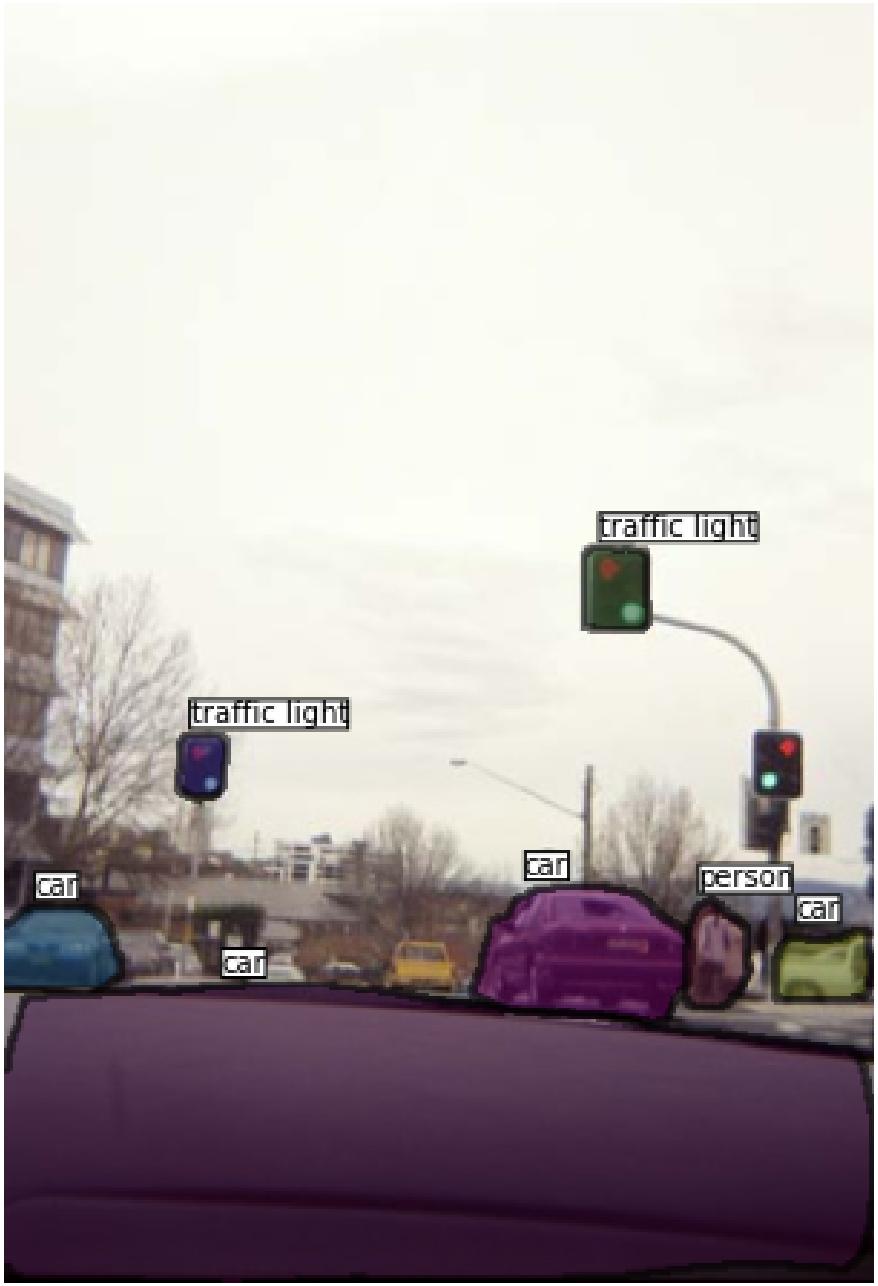
Results

Results


broccoli
broccoli
carrot
carrot
broccoli
broccoli
carrot
broccoli
carrot
broccoli
carrot
broccoli

Mistakes


person
baseball glove
tennis racket
person


traffic light
traffic light
car
car
car
person
car

Results

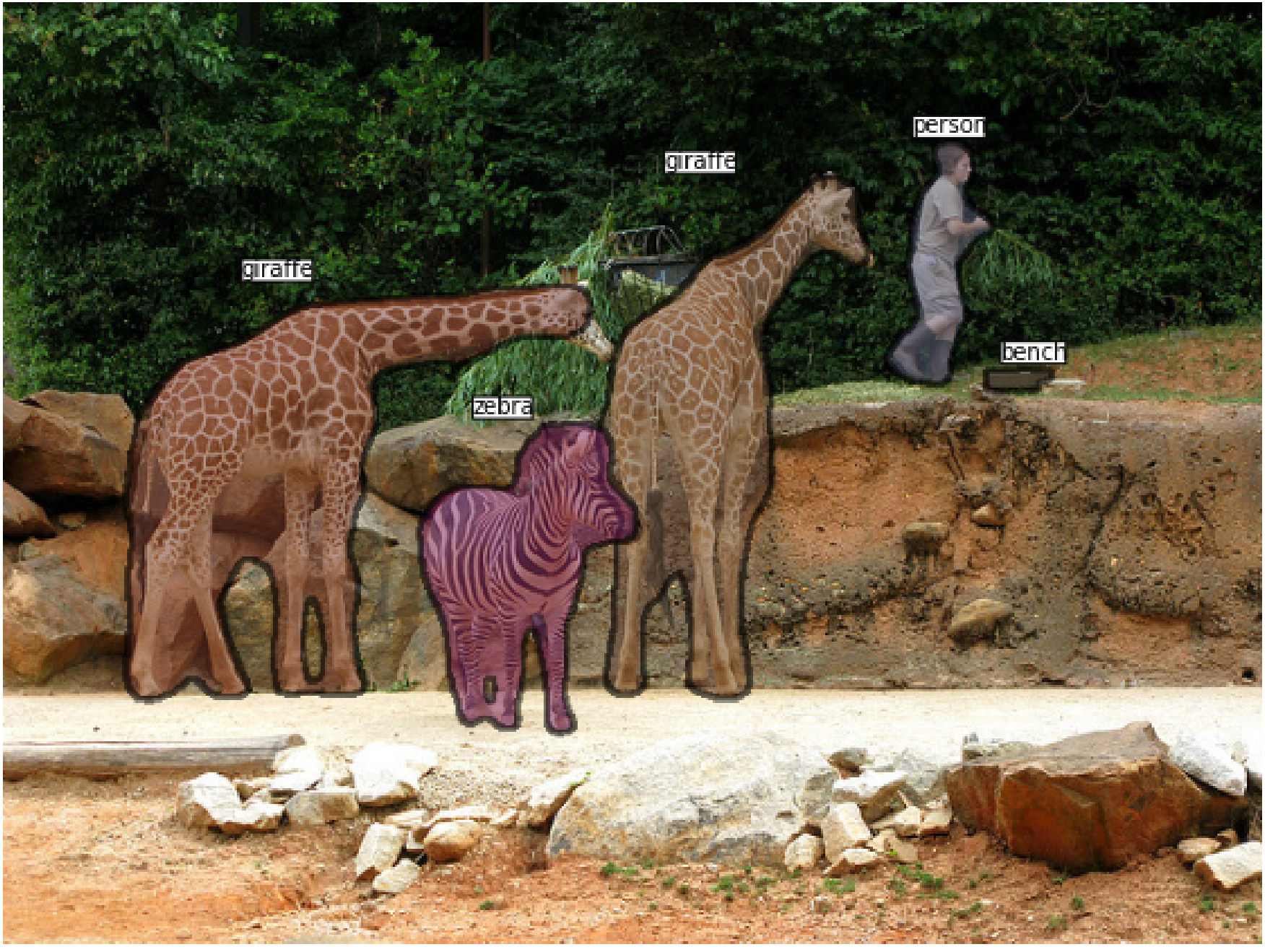
Results


sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep

Results

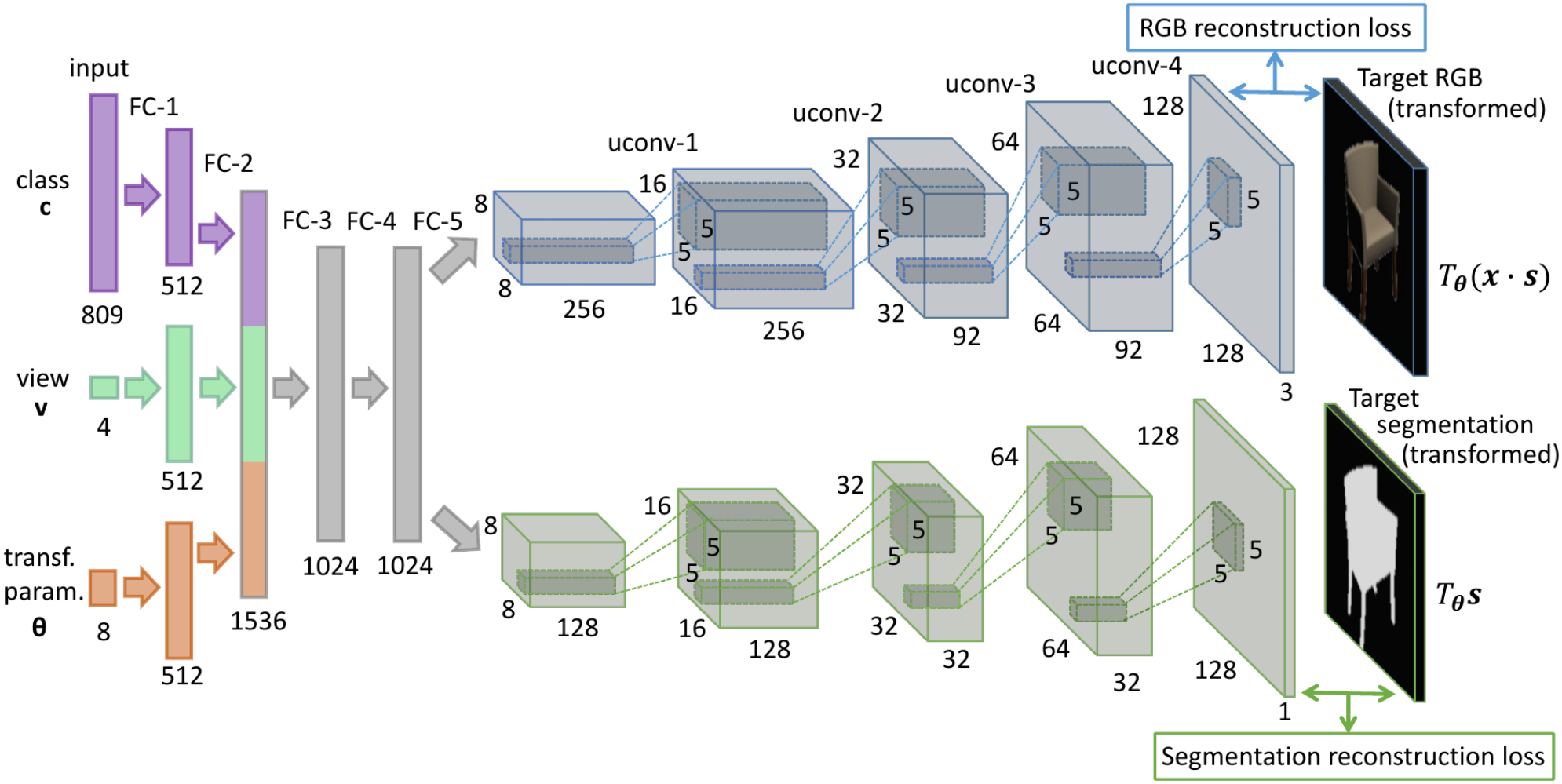

donut
sandwich
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut

Results



Supervised ConvNets that Draw Pictures

- Using ConvNets to Produce Images
- [Dosovitskyi et al. Arxiv:1411:5928

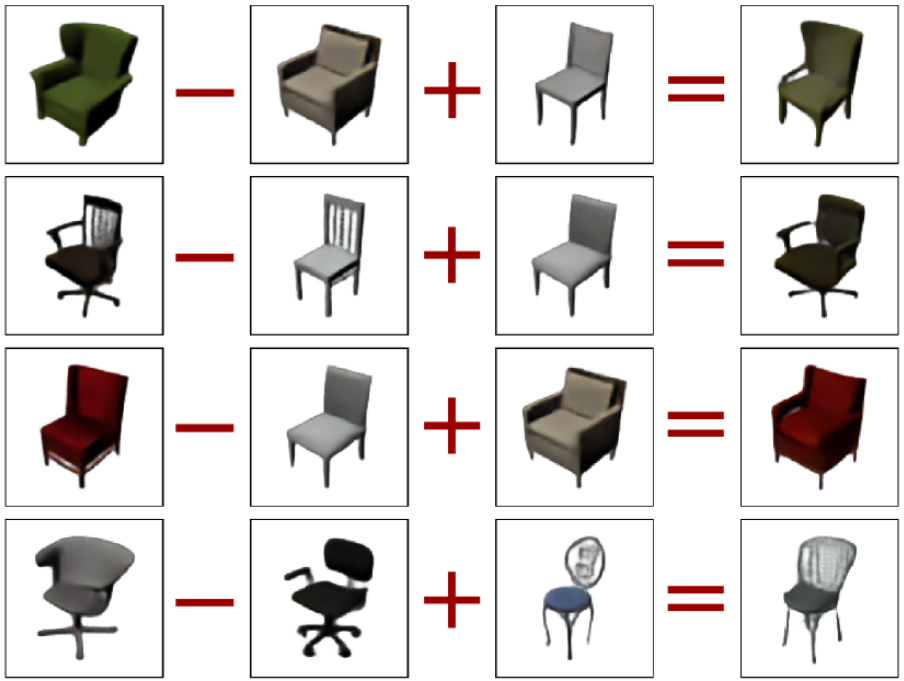


Supervised ConvNets that Draw Pictures

Y LeCun

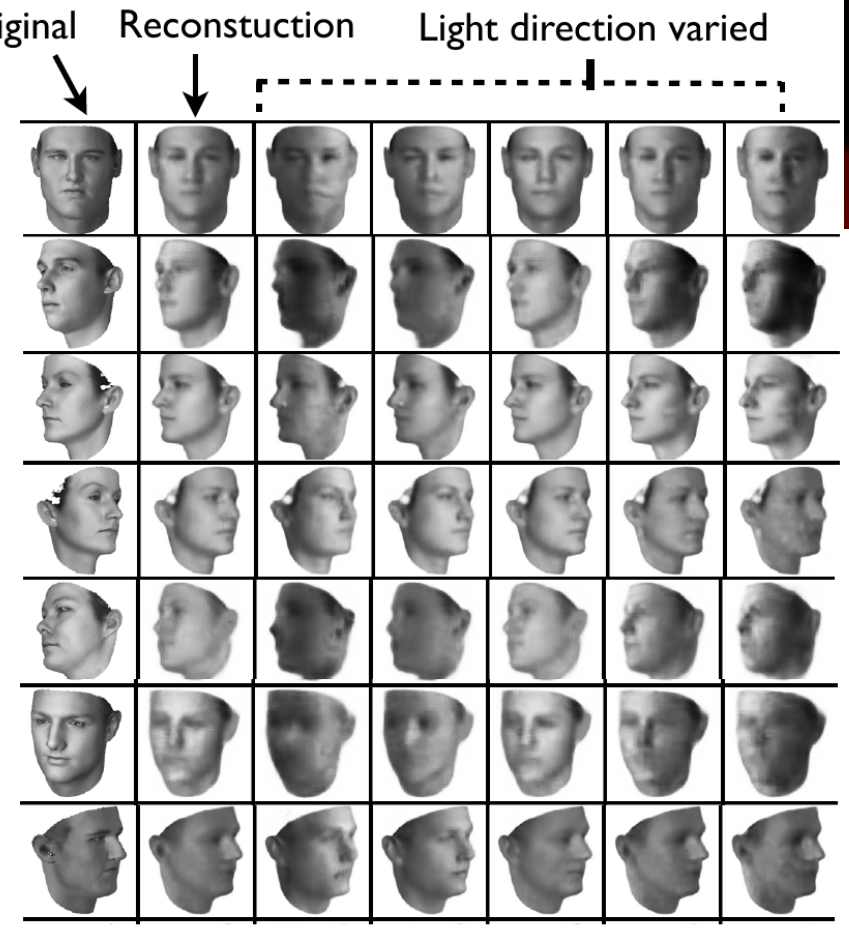
Generating Chairs

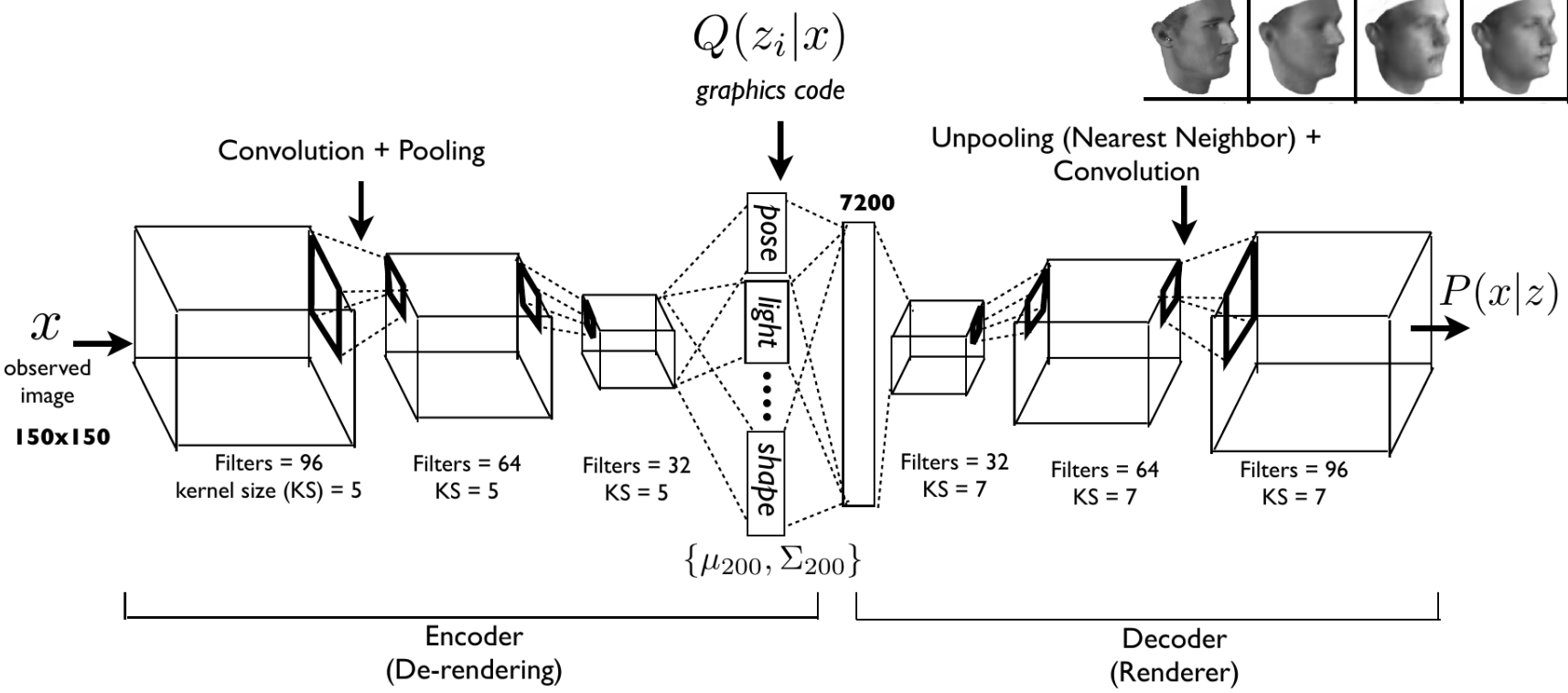
Chair Arithmetic in Feature Space



Convolutional Encoder-Decoder

- Generating Faces
- [Kulkarni et al. Arxiv:1503:03167]

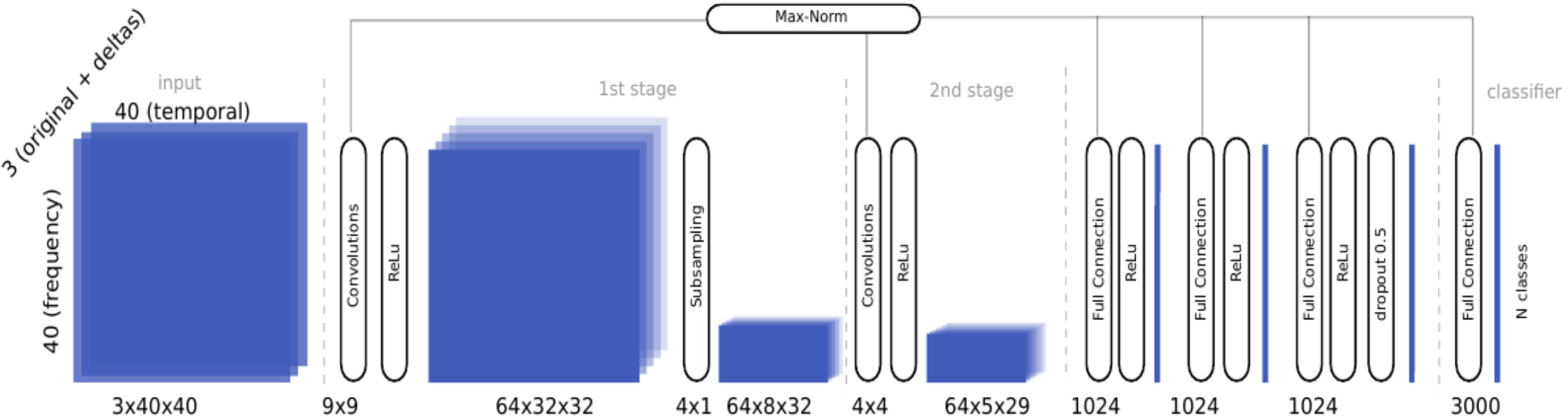





Speech Recognition
With
ConvNets

Speech Recognition with Convolutional Nets (NYU/IBM)

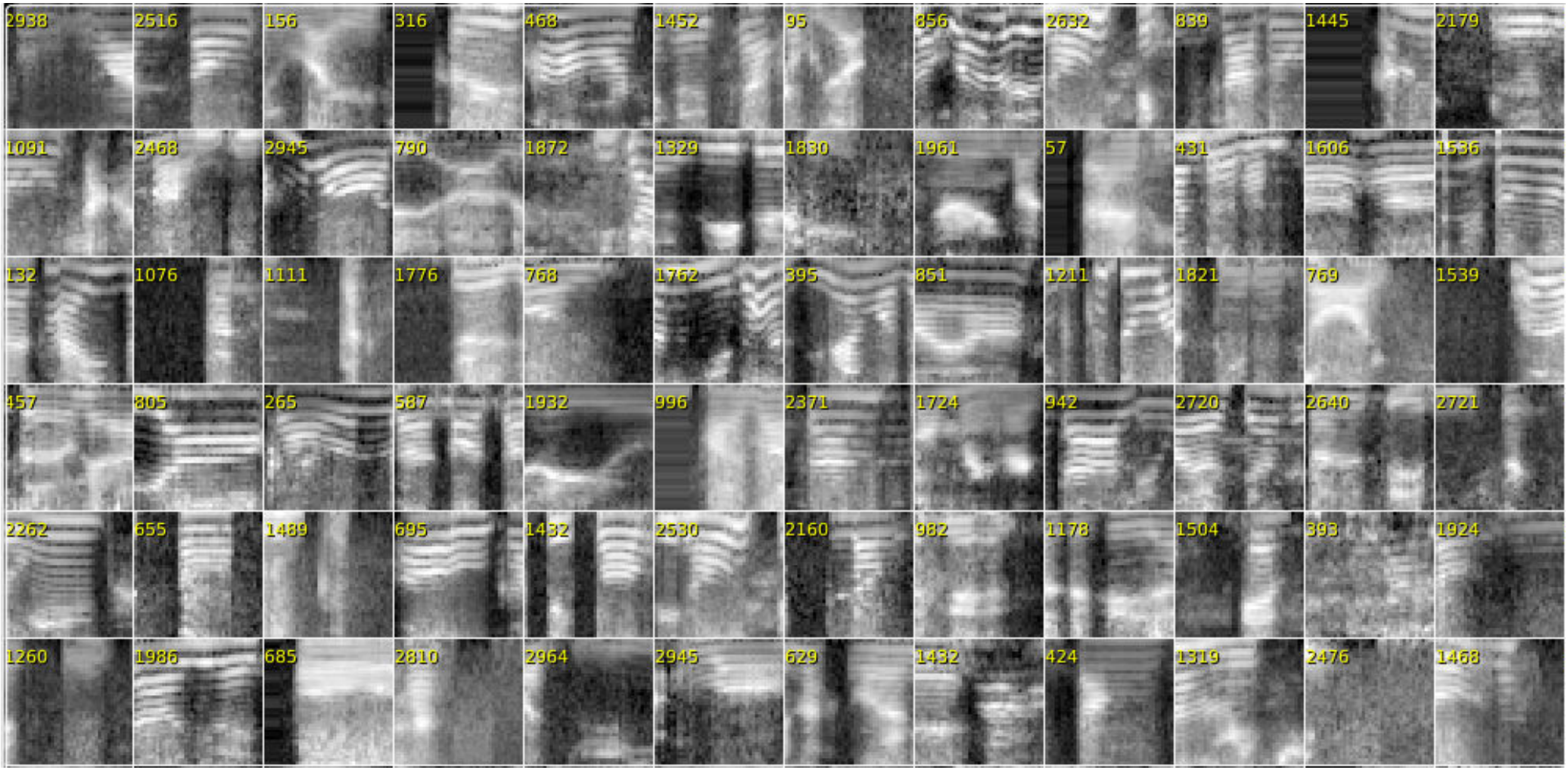
Y LeCun



- Acoustic Model: ConvNet with 7 layers. 54.4 million parameters.
- Classifies acoustic signal into 3000 context-dependent subphones categories
- ReLU units + dropout for last layers
- Trained on GPU. 4 days of training

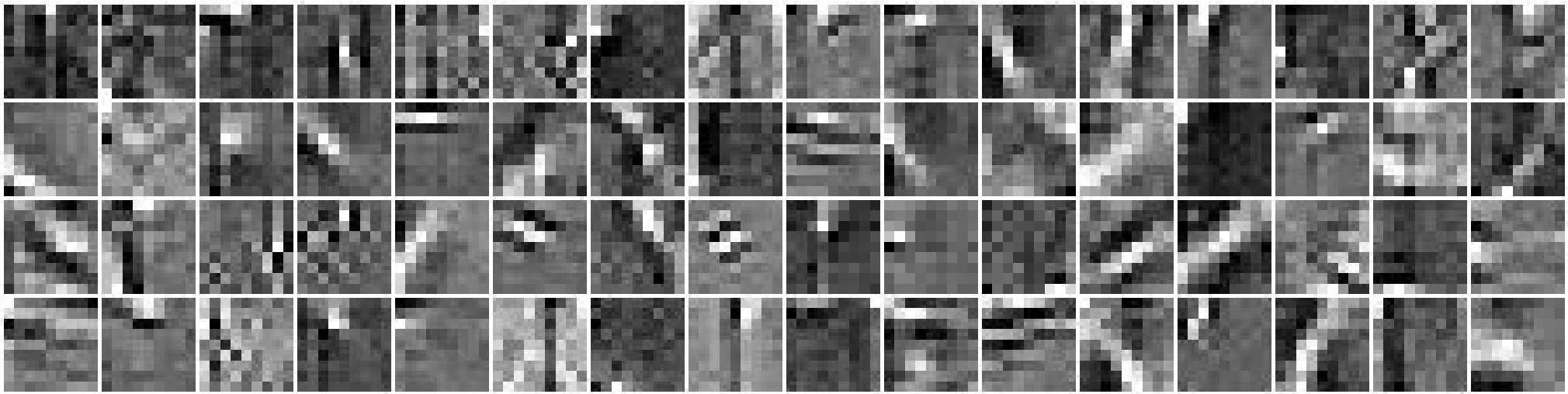
Speech Recognition with Convolutional Nets (NYU/IBM)

- Training samples.
  - 40 MEL-frequency Cepstral Coefficients
  - Window: 40 frames, 10ms each



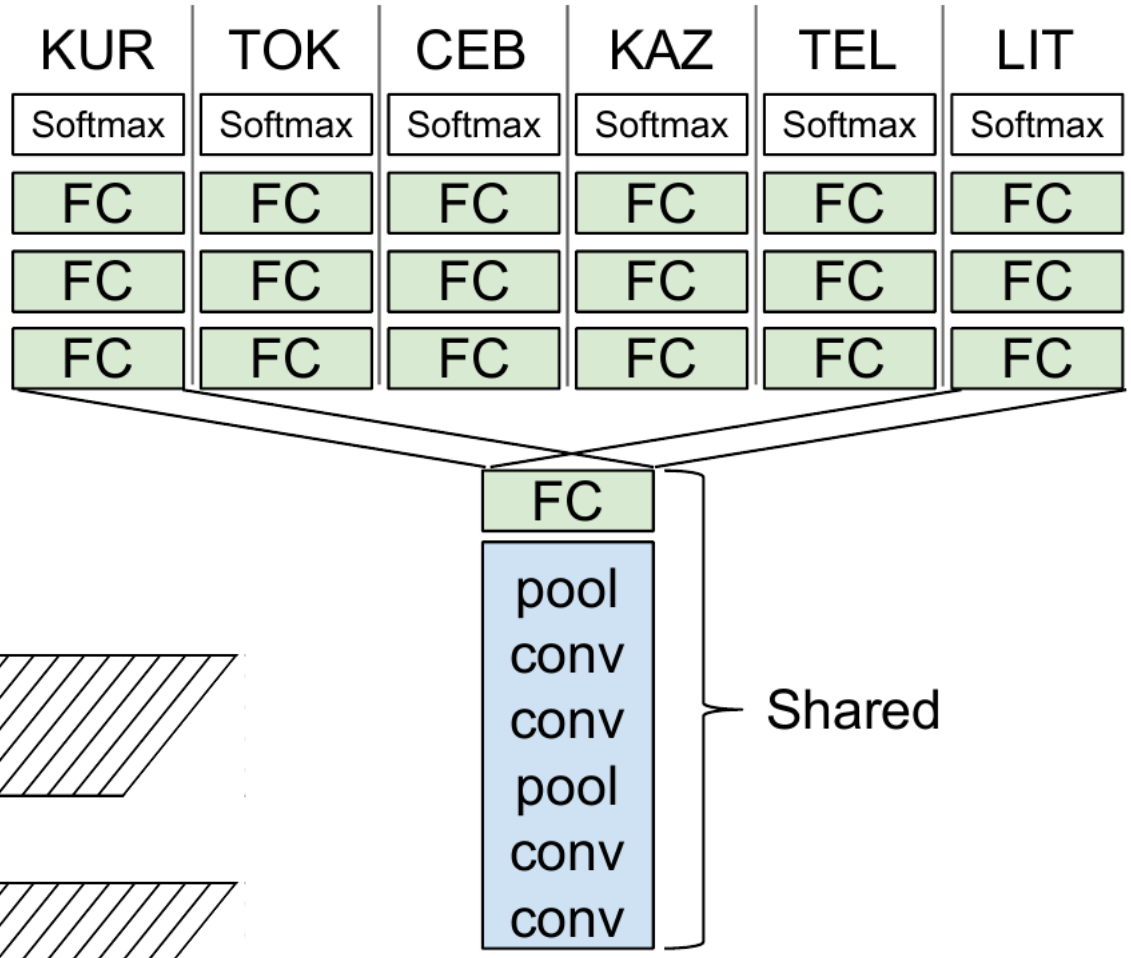
Speech Recognition with Convolutional Nets (NYU/IBM)

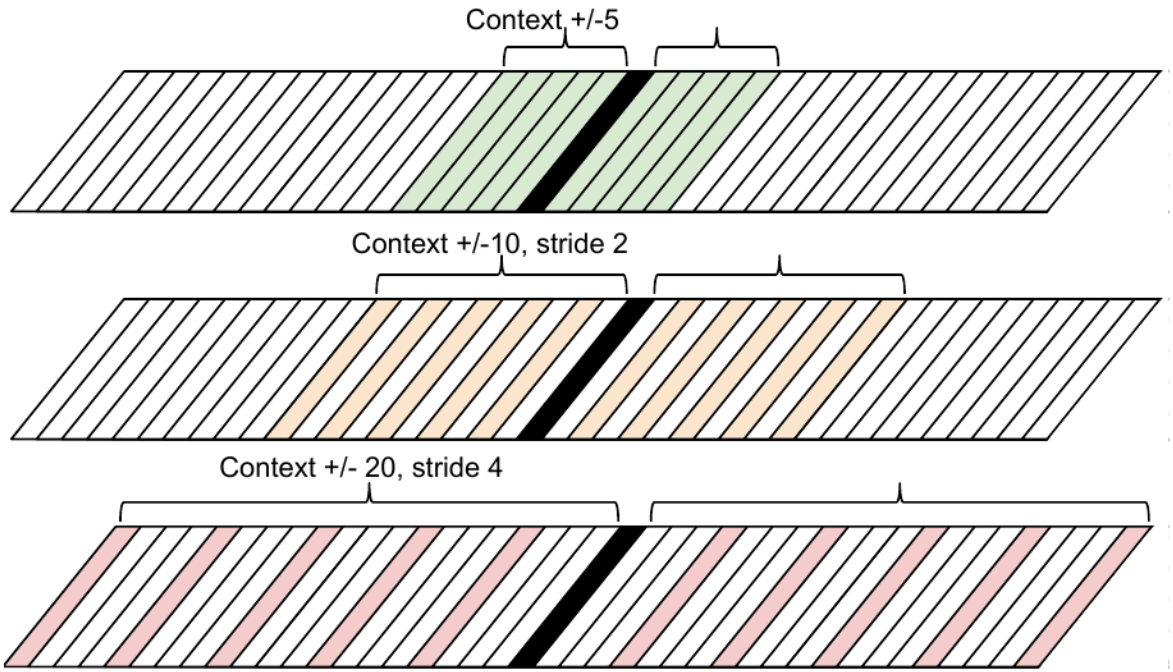
- Convolution Kernels at Layer 1:
  - 64 kernels of size 9x9



Speech Recognition with Convolutional Nets (NYU/IBM)

- Multilingual recognizer
- Multiscale input
  - Large context window






ConvNets are Everywhere
(or soon will be)

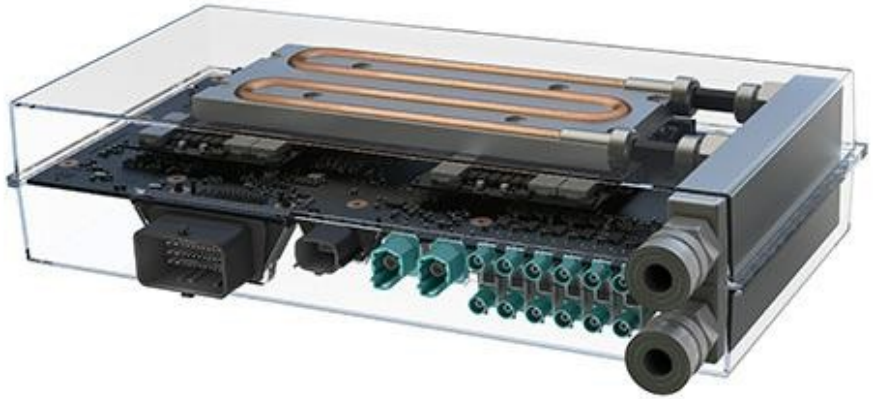
NVIDIA: ConvNet-Based Driver Assistance

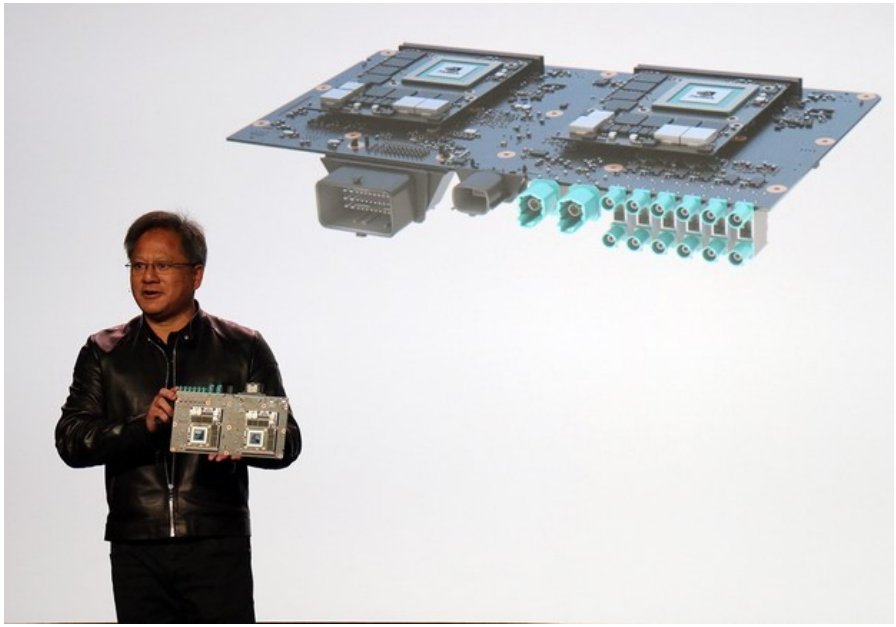
Y LeCun

Drive-PX2: Open Platform for Driver Assistance

Embedded Super-Computer: 42 TOPS

- (=150 Macbook Pros)



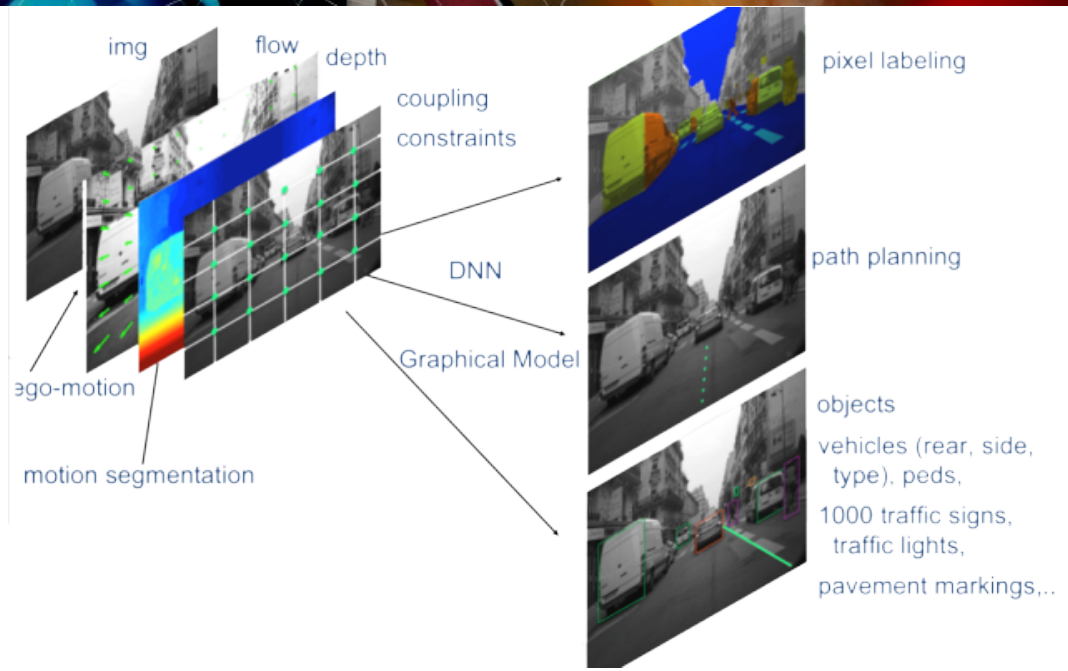


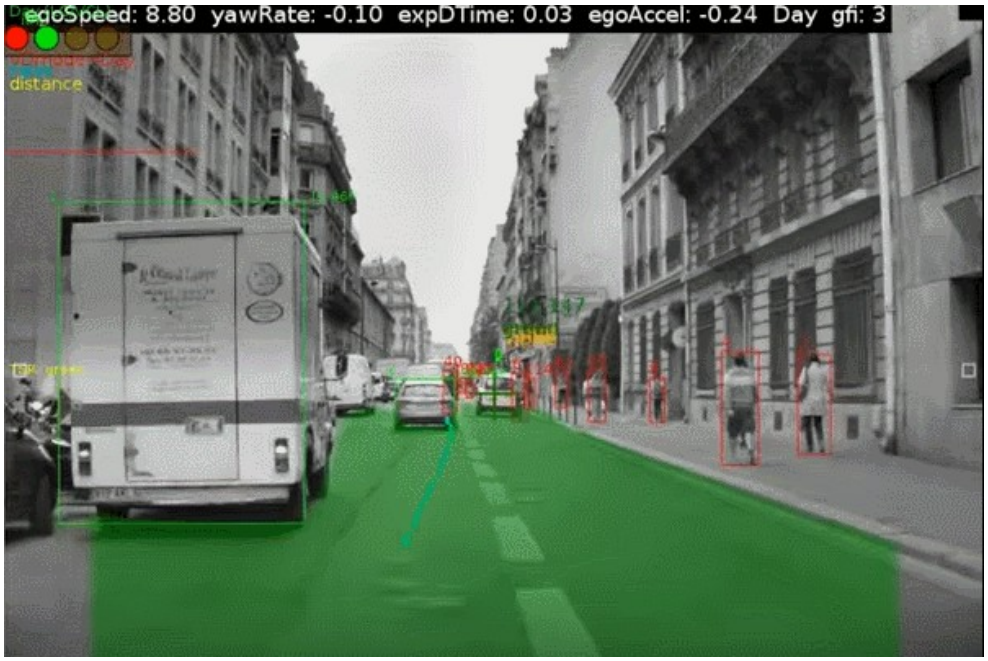
MobilEye: ConvNet-Based Driver Assistance

Y LeCun

Deployed in the latest

Tesla Model S and Model X



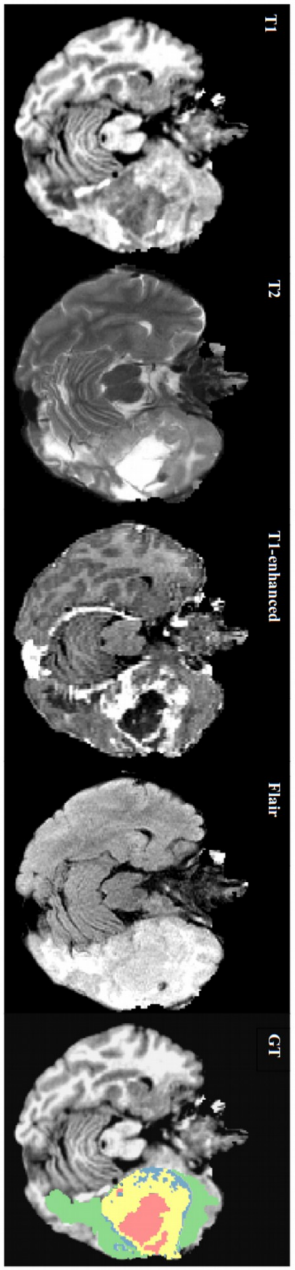


ConvNet in Connectomics [Jain, Turaga, Seung 2007]

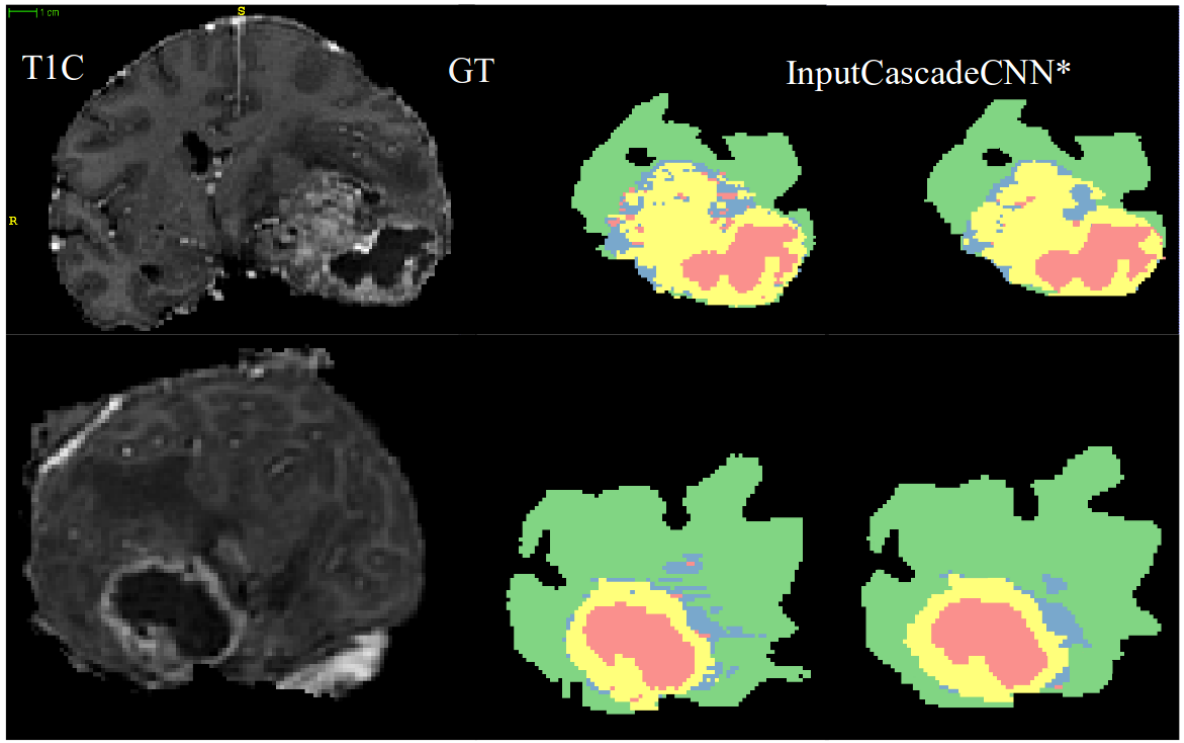
- 3D ConvNet
  - Volumetric
  - Images
- Each voxel labeled as “membrane” or “non-membrane using a 7x7x7 voxel neighborhood
- Has become a standard method in connectomics

VIDEO

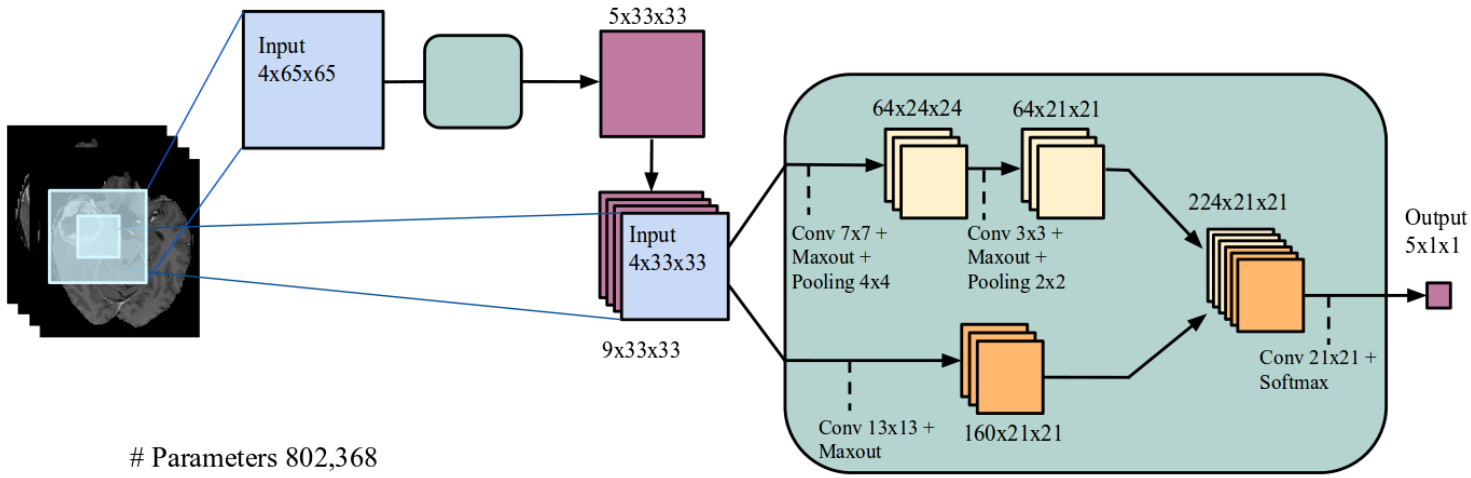
Brain Tumor Detection



- [Havaei et al. 2015]
  - Arxiv:1505.03540
- InputCascadeCNN architecture
  - 802,368 parameters
- Trained on 30 patients.
- State of the art results on BRAT2013



edema, enhanced tumor, necrosis, non-enhanced tumor.



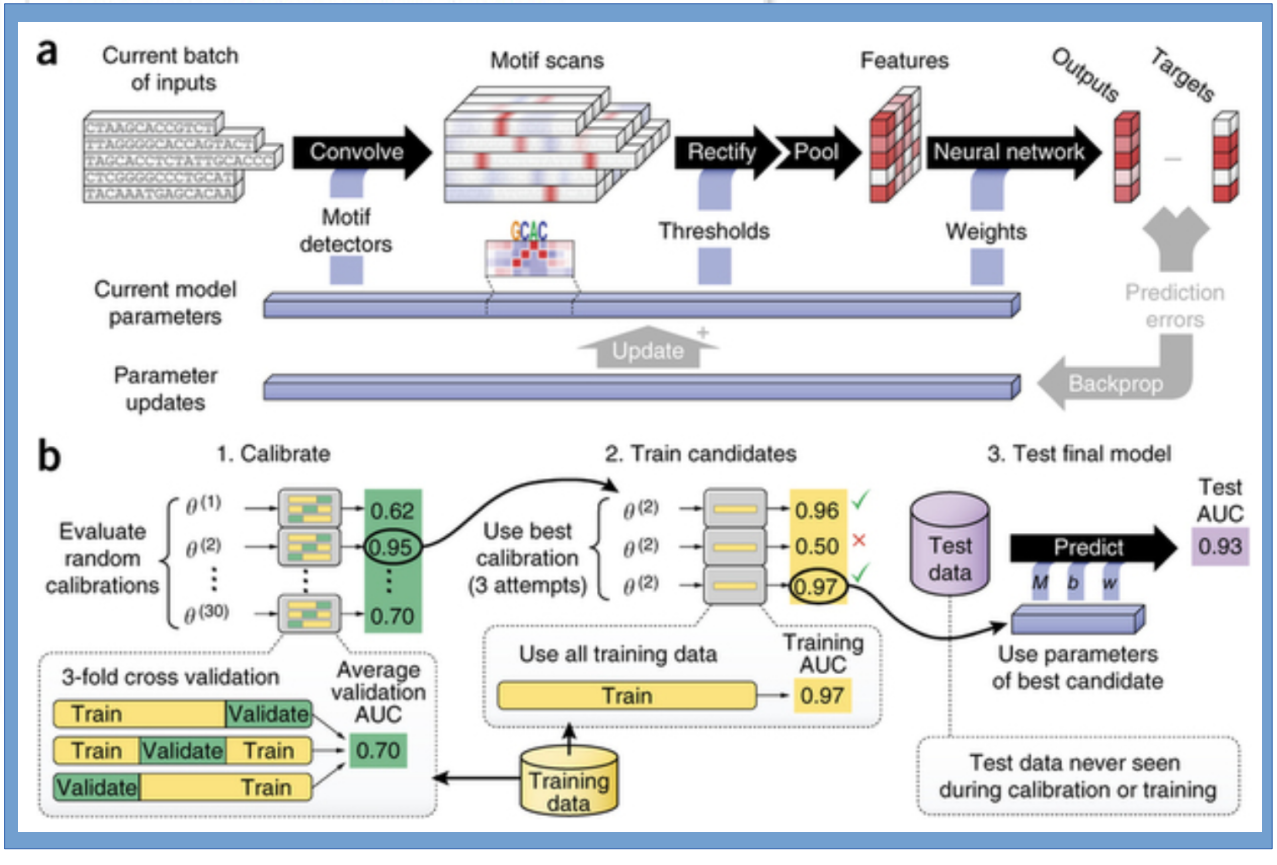
Parameters 802,368

Predicting DNA/RNA – Protein Binding with ConvNets

Y LeCun

"Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning" by B Alipanahi, A Delong, M Weirauch, B Frey,

Nature Biotech, July 2015.



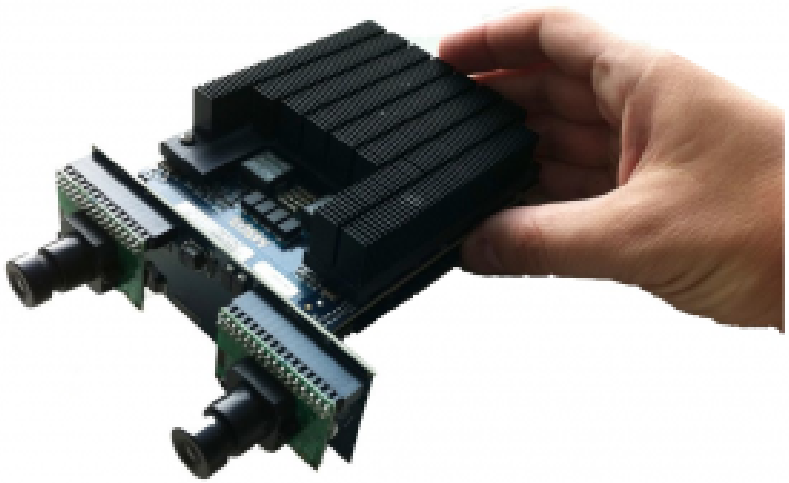
Deep Learning is Everywhere (ConvNets are Everywhere)

- **Lots of applications at Facebook, Google, Microsoft, Baidu, Twitter, IBM…**
  - Image recognition for photo collection search
  - Image/Video Content filtering: spam, nudity, violence.
  - Search, Newsfeed ranking
- **People upload 800 million photos on Facebook every day**
  - (2 billion photos per day if we count Instagram, Messenger and Whatsapp)
- **Each photo on Facebook goes through two ConvNets within 2 seconds**
  - One for image recognition/tagging
  - One for face recognition (not activated in Europe).
- **Soon ConvNets will really be everywhere:**
  - self-driving cars, medical imaging, augemnted reality, mobile devices, smart cameras, robots, toys…..

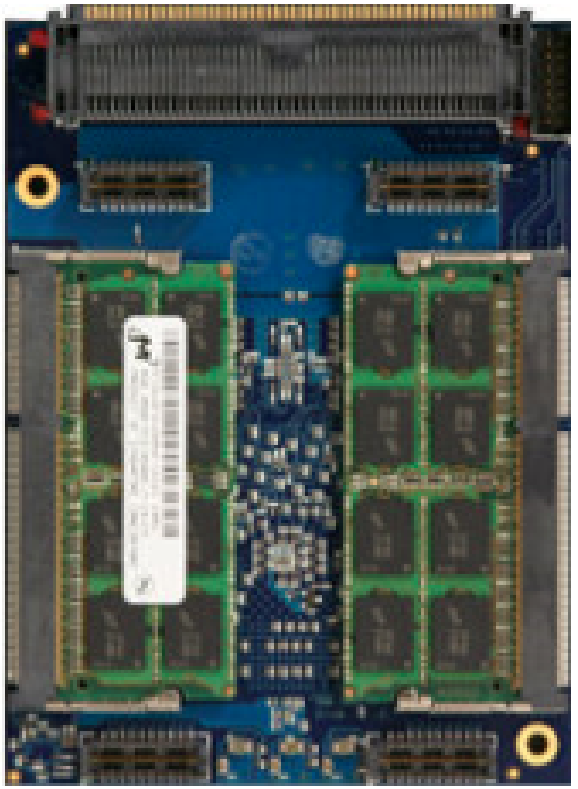

ConvNet
Hardware

NeuFlow architecture (NYU + Purdue)

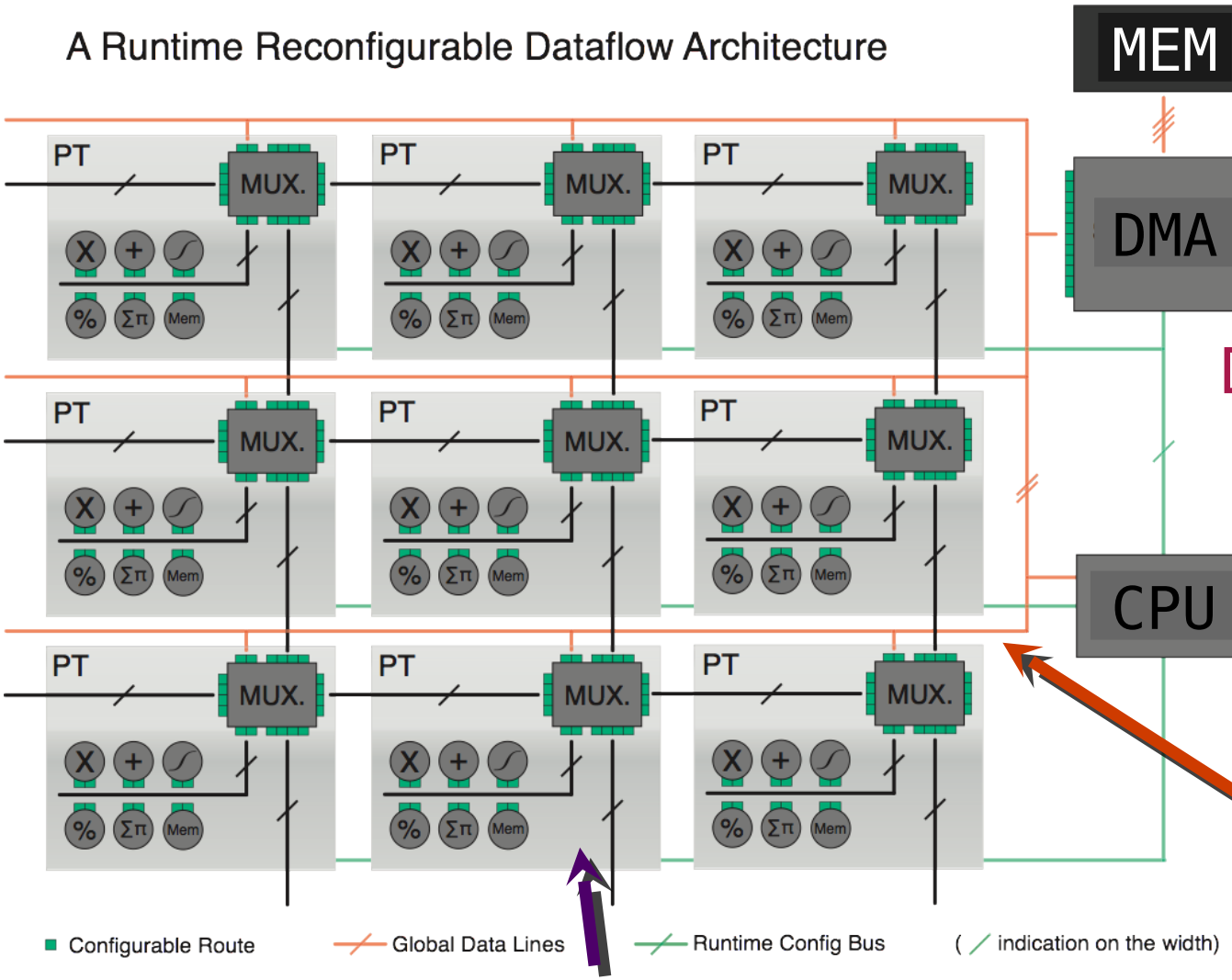
- Collaboration NYU-Purdue: Eugenio Culurciello's e-Lab.
- Running on Picocomputing 8x10cm high-performance FPGA board
  - Virtex 6 LX240T: 680 MAC units, 20 neuflow tiles
- Full scene labeling at 20 frames/sec (50ms/frame) at 320x240



board with Virtex-6



NewFlow: Architecture



MEM

DMA

Multi-port memory controller (DMA)

[x12 on a V6 LX240T]

CPU

RISC CPU, to reconfigure tiles and data paths, at runtime

global network-on-chip to allow fast reconfiguration

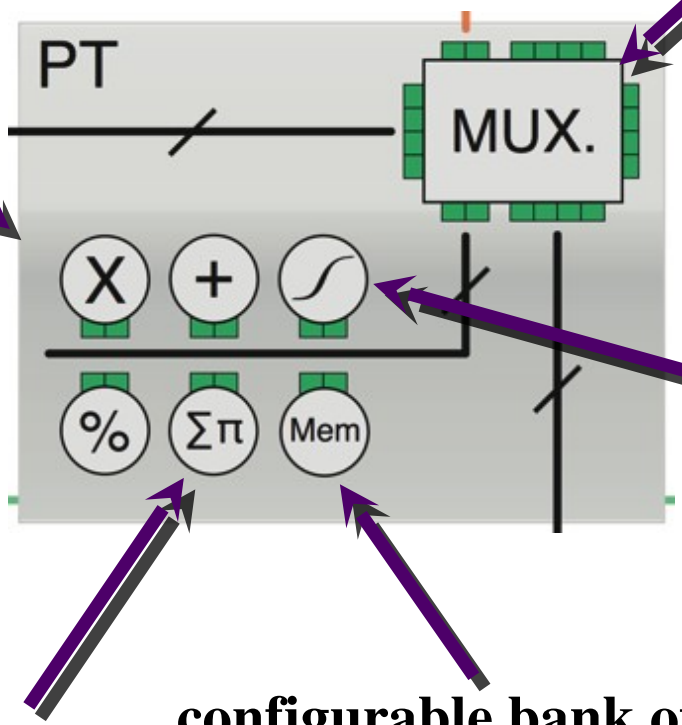
grid of passive processing tiles (PTs)

[x20 on a Virtex6 LX240T]

NewFlow: Processing Tile Architecture

Term-by:-term streaming operators (MUL,DIV,ADD,SUB,MAX)

[x8,2 per tile]



configurable router, to stream data in and out of the tile, to neighbors or DMA ports

[x20]

configurable piece-wise linear or quadratic mapper

[x4]

full 1/2D parallel convolver with 100 MAC units

[x4]

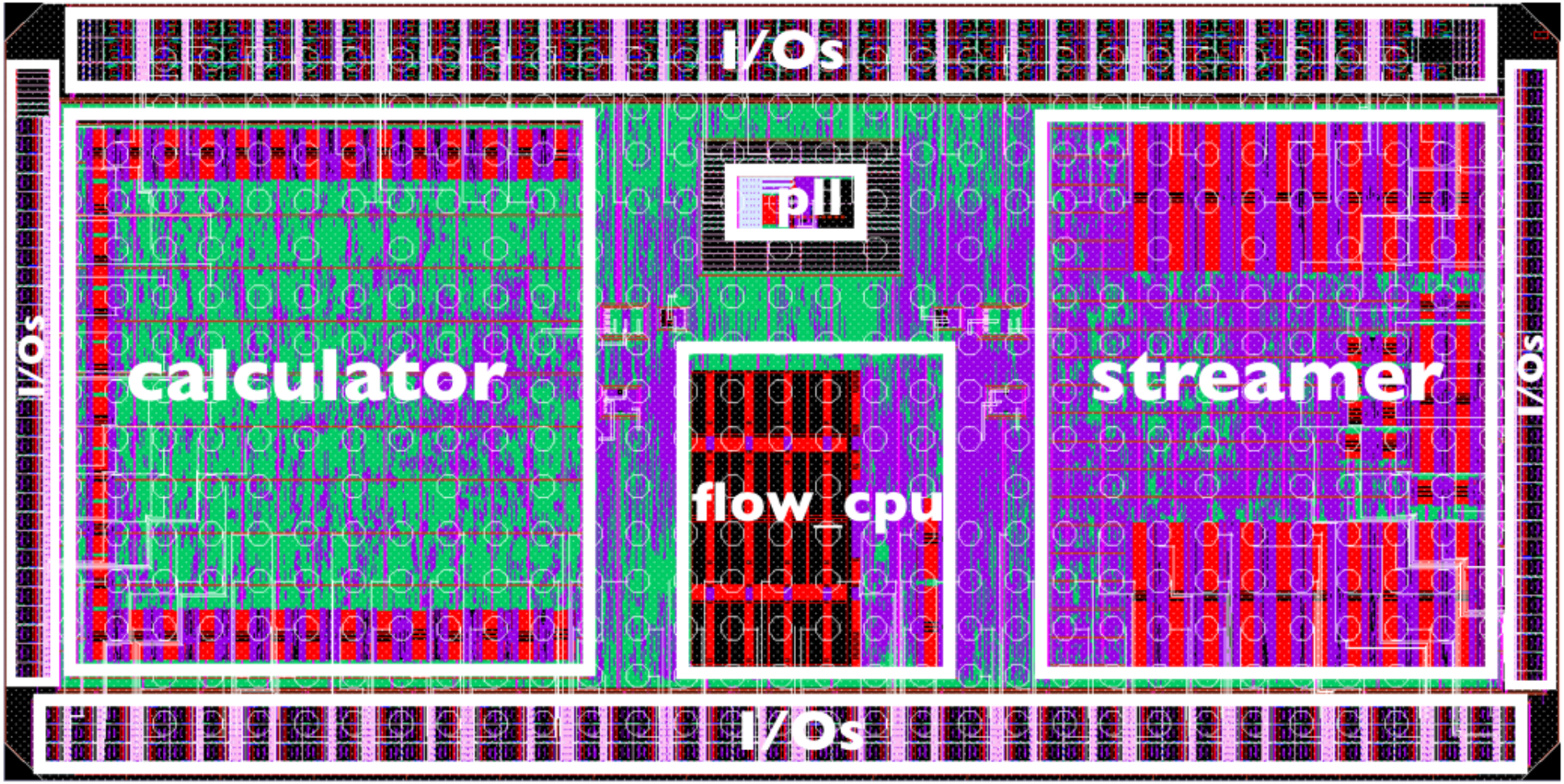
configurable bank of FIFOs , for stream buffering, up to 10kB per PT

[x8]

[Virtex6 LX240T]

NewFlow ASIC: 2.5x5 mm, 45nm, 0.6Watts, >300GOPS

- Collaboration Purdue-NYU: Eugenio Culurciello's e-Lab
- Suitable for vision-enabled embedded and mobile devices
- (but the fabrication was botched...)



[Pham, Jelaca, Farabet, Martini, LeCun, Culurciello 2012]

ConvNet Hardware Today

- **Training is all done on GPUs (on NVIDIA GPUs)**
  - 4 or 8 GPU cards per node. Many nodes per rack. Many racks.
  - 5 Tflops per card (2014)
  - Training needs performance, programmability, flexibility, accuracy.
  - Power consumption and space is not that important.
- **Many large hardware companies are developing ConvNet accelerators**
  - NVIDIA: evolving from GPU. Embedded applications
    - NVIDIA is betting the farm on deep learning.
  - Intel: ConvNet Fixed Function core. 10X over standard core
  - Mobileye: ConvNet chip for automotive
  - Orcam: low-power ConvNet chip for the visually impaired
  - Qualcomm, Samsung: ConvNet IP for mobile devices
- **Lots of startups are getting into the field**
  - Teradeep, derived from NeuFlow
  - Many others...

Deep Learning Applications Today = ConvNets

- **Practical/deployed applications of deep learning:**
  - Almost exclusively ConvNet trained with Backprop
  - A few recurrent nets used experimentally
  - Applications: image recognition, face recognition, video understanding, speech recognition, natural language understanding.
- **Training**
  - Backprop running on racks of GPUs
- **Production**
  - ConvNets on CPUs, with reduced-precision arithmetics
  - Some are moving to GPUs or FPGA.
- **Scale**
  - People upload 800 million images on Facebook every day
  - All of them go through two ConvNets within 2 seconds
    - One for image tagging, one for face recognition.